Evolutionary Pattern of Gene Homogenization between Primate-Specific Paralogs after Human and Macaque Speciation Using the 4-2-4 Method

Kiyoshi Ezawa,[1,2] Kazuho Ikeo,[2,3] Takashi Gojobori,[2,3] and Naruya Saitou*[,1]

[1]Division of Population Genetics, National Institute of Genetics, Mishima, Japan

[2]Human Genome Network Project, National Institute of Genetics, Mishima, Japan

[3]DNA Data Analysis Laboratory, National Institute of Genetics, Mishima, Japan

***Corresponding author:** E-mail: saitounr@lab.nig.ac.jp.

Associate editor: Anne Stone

Research article

Abstract

Homogenization of duplicated genes is an important factor for gene family evolution. In the previous study, we developed a method, named 4-2-4 here, to detect partial homogenization with high sensitivity and high specificity using quartets. A quartet is a set of four genes generated by a duplication event and the subsequent speciation of two closely related species. We searched the human and macaque genomes and found 430 nonredundant quartets, which correspond to primate-specific paralogs. The prevalence of homogenization in these quartets was 10.0% (43/430), which was ca. one-third of that (29.8% = 206/691) in the rodent-specific nonredundant quartets obtained through comparison of mouse and rat genomes. Part of this difference comes from the fact that primate paralogs tend to be more remotely located to each other than rodent paralogs, and the remainder may be explained by the inherent difference in the neutral evolutionary rate between the primate and rodent lineages. A statistical analysis taking account of the effects of false negatives uncovered negative correlations between sequence divergence and homogenization prevalence both in primates and rodents. Further statistical analyses controlling for false-negative rates and sequence divergences revealed two characteristics shared by the primate and rodent paralogs; 1) significant negative correlations of the homogenization prevalence with physical distances, and 2) no significant correlation between the prevalence and relative transcriptional orientations. Patterns of the homogenization in the genomic alignments of human–macaque quartets indicate that gene conversion, rather than unequal crossing-over, is the major cause of the homogenization.

Key words: gene conversion, duplicated genes, human, macaque, 4-2-4 method, genome-wide analysis.

Introduction

When gene duplication occurs, duplicated copies start to accumulate mutations from the ancestral sequence. This sequence divergence process may be hampered by homogenization, or retainment of high sequence similarities with their duplicate partners (Brown et al. 1972; Arnheim 1983; Nei and Rooney 2005). Two major mechanisms have been proposed so far to explain the homogenization; one is unequal crossing-over (Smith 1976; Ohta 1976) and the other is gene conversion (Jeffreys 1979; Slightom et al. 1980; Ohta 1985). Unequal crossing-over takes place between nonequivalent but homologous regions of a pair of sister chromatids or homologous chromosomes, as is well documented on rRNA genes (Eickbush and Eickbush 2007). When occurring in an intergenic region, unequal crossing-over only changes the copy numbers of tandemly duplicated genes. When occurring in an intragenic region, however, it can also create a chimera of two mutually homologous genes, which exhibits complementary homogenization with its parent sequences. Gene conversion, also known as nonreciprocal recombination, is a process where a tract of DNA overwrites a homologous one (Petes and Hill 1988; chapter 11 of Li 1997). Gene conversion can be classified into intralocus and interlocus types. Interlocus gene conversion between duplicate genes causes homogenization of relatively short regions, as observed between red and green opsin genes in catarrhine primates (Nathans et al. 1986; Ibbotson et al. 1992; Balding et al. 1992).

If the homogenization is relatively sporadic and regionally limited, it causes conflicts between locally inferred phylogenetic relationships (Scott et al. 1984; Kawamura et al. 1992; Shyue et al. 1994; Zhou and Li 1996; Cheung et al. 1999; Kitano and Saitou 1999; Winter and Ponting 2005). If it occurs frequently and/or extensively, homogenization can lead to erroneous phylogenetic inferences or misestimation of duplication dates, thus can confound estimation of evolutionary history of gene families (Slightom et al. 1985; Gao and Innan 2004; Schienman et al. 2006). It is therefore crucial to grasp the pattern of genome-wide prevalence of homogenization, namely the proportion of duplicate pairs that underwent homogenization, as well as properties of gene pairs that enhance or reduce the frequency of homogenization.

We are interested in the prevalence of interlocus homogenization in the evolutionary history of the human genome in the present study. Analyses of homogenization for

the human genome have been conducted by Jackson et al. (2005), and recently by Benovoy and Drouin (2009) and McGrath et al. (2009). Jackson et al. (2005) gathered 24 duplicon families with sequence divergence of at most 4%, and applied a "quartet method" they developed to the multiple alignments of the duplicon families. Their method is different from our "4-2-4" method in that their quartets consist only of duplicates of a single species, whereas our quartets consist of duplicates of two closely related species. They detected homogenized regions in at least 5% of the sequence alignments whose total length was >8 Mb. Benovoy and Drouin (2009) examined 55,050 pairs of human duplicate genes showing more than 60% sequence identity and detected homogenization in only 483 pairs (0.88%) using GENECONV. McGrath et al. (2009) applied GENECONV to the sets of species-specific duplicate genes in human, macaque, mouse, and rat genomes. They estimated homogenization rate as 12.5% for human gene pairs and 14.6% for mouse gene pairs. These three studies showed an enormous variation (from <1% to ~13%) in the prevalence of homogenization between human duplicates.

Benovoy and Drouin (2009) and McGrath et al. (2009) also tried to identify the properties of duplicate pairs that enhance the occurrence probability of homogenization. Whereas Benovoy and Drouin (2009) found a negative correlation between susceptibility to homogenization and sequence divergence, McGrath et al. (2009) failed to find such a correlation. The latter group attributed this failure to the high false-negative rate of GENECONV applied to a pair of highly similar genes (table S1 of their paper). Both groups concluded that the homogenization susceptibility does not significantly depend on the physical distance between duplicates after controlling for the dependence on the sequence identity. This, however, seems inconsistent with the study on mouse/rat quartets by Ezawa et al. (2006), who found that homogenization susceptibility negatively correlated with the physical distance.

What caused such inconsistencies among different studies? Inconsistencies are mainly due to the difference in the conditions to collect duplicate pairs and the difference in the homogenization detection methods and criteria. First, the enormous differences in the homogenization rate (<1% vs. 13% in human) is due to the fact that the data set of Benovoy and Drouin (2009) mainly consists of sequence pairs whose members are so divergent with each other that they rarely underwent recent gene conversion. GENECONV cannot efficiently detect ancient homogenization whose tracts subsequently accumulated multiple substitutions. It is misleading to include such divergent pairs when estimating the impacts of homogenization, although such pairs may be important when observing how the sequence divergence reduces the homogenization rate. The presence/absence of a correlation between the "homogenization prevalence" and the sequence similarity must partly be due to this inclusion/exclusion of highly divergent pairs in their studies. The high false-negative rate of GENECONV on highly similar pairs is another problem as mentioned above.

Another inconsistency among previous studies is the correlation between the homogenization prevalence and the physical distance. Ezawa et al. (2006) analyzed this correlation, but their study had two problems: 1) it just appealed to visual images and did not conduct statistical tests to quantify the significance of the correlation; 2) it did not take account of the dependence on the sequence divergence. Although Benovoy and Drouin (2009) and McGrath et al. (2009) did some statistical tests, their studies also had problems. The most serious one is that both studies included highly similar gene pairs on which GENECONV show extremely poor true-positive rate (e.g., <20% of homogenization detection as shown in table S1 of McGrath et al. 2009). This means that their analyses could have called gene pairs as "negative" for homogenization when they had actually undergone homogenization so intensively that they have kept high sequence similarities. Assuming that such gene pairs tend to have small physical distances, it would not be surprising if they missed a negative correlation that actually exists between the homogenization rate and the physical distance.

We would like to uncover the real biological properties of human paralog homogenization that might have been hidden by artifacts in the previous studies. We thus analyzed human and macaque genome data under a setting in which artificial effects are well controlled. For comparison, we also reanalyzed mouse and rat genome data used by Ezawa et al. (2006). Main features of our well-controlled setting are the following (for details, see section A of supplementary materials and methods, Supplementary Material online):

- (i) Use of quartets for the homogenization detection (fig. 1);
- (ii) Use of moderately diverged duplicate pairs;
- (iii) Use of simulated quartets to estimate false-negative rates;
- (iv) Use of the proportion of positive quartets rather than that of "positive gene pairs";
- (v) Use of logistic regression methods to estimate the statistical significance of correlations.

Under such a well-controlled setting, we were able to quantify the degree of statistical significance of the dependence on the properties of duplicate pairs. This is the first study that took account of the effects of artifacts such as false-negative rates, and the results shown in this paper are therefore regarded as reliably reflecting the biological trends of the homogenization susceptibility. This well-controlled setting combined with the logistic regression analysis also enabled the comparison of homogenization prevalence between primates and rodents after controlling for various factors. We believe that this study will serve as a sound cornerstone that the future genome-wide analyses on interlocus homogenization will be based on.

Another important and outstanding issue is the major mechanism that caused the observed homogenization of duplicated genes in the genome. Ezawa et al. (2006) assumed that gene conversion should be the dominant

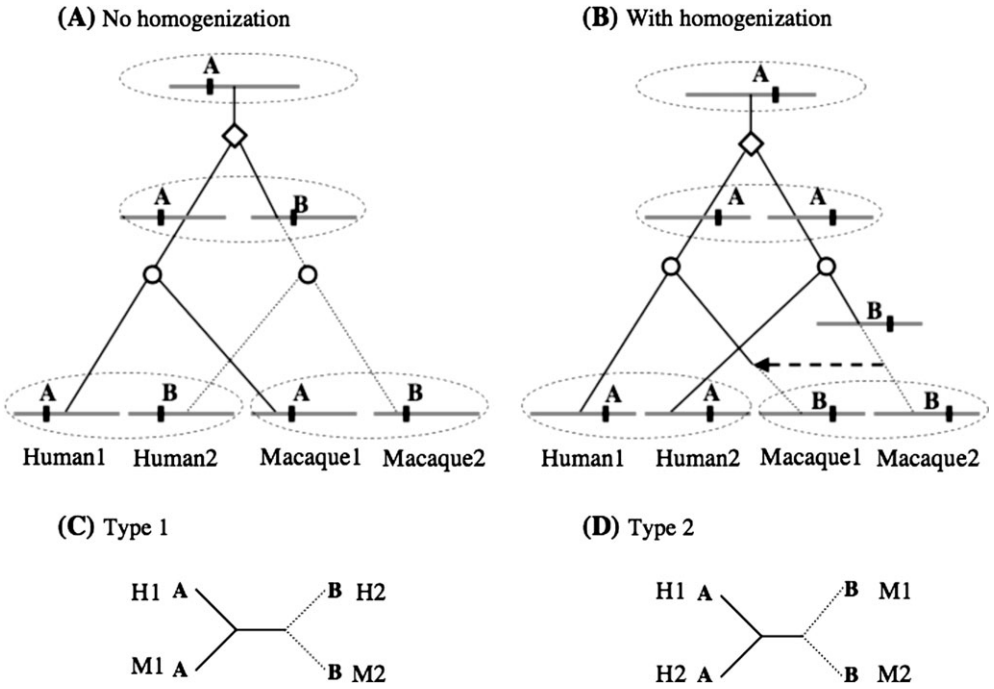

FIG. 1. Evolution of a duplicate pair with and without homogenization (*A*, *B*) and resulting informative sites in a quartet (*C*, *D*). This figure illustrates the evolution of a gene in the common ancestor of two species (human and macaque as an example) that underwent duplication before speciation, generating a quartet of genes, Human1, Human2, Macaque1, and Macaque2. Panel (*A*) shows the typical evolution of a nucleotide in a quartet, where an ancestral nucleotide A is substituted with B after duplication and before speciation. This results in a type I informative site, which clusters orthologous genes together (*C*). Panel (*B*) illustrates the evolution of a base involved in homogenization, where a substitution of the nucleotide A with B occurs after the speciation, then homogenization propagates the base B to the intraspecies paralog (the horizontal dashed arrow). The resulting informative site is type II, which clusters intraspecies paralogs (*D*). NOTE.—A thick gray horizontal bar denotes a gene, and a short black rectangle on the bar denotes the nucleotide site in question. In each tree, branches in solid lines give trajectories of the nucleotide A, and branches in dotted lines give trajectories of the nucleotide B. An open diamond and an open circle represent a prespeciation duplication event and a speciation event, respectively. An ellipse in a dashed line indicates that the genes are in the same genome. A horizontal arrow in a dashed line denotes homogenization of a nucleotide site. In panels (*C*) and (*D*), the symbols "H1," "H2," "M1," and "M2" stand for Human1, Human2, Macaque1, and Macaque2, respectively.

mechanism and referred to the homogenization as "gene conversion.". In some cases, however, homogenization seems dominated by unequal crossing-over (Eickbush and Eickbush 2007), which also seems to be one of the driving forces of the "birth-and-death" model of gene family evolution (Nei and Rooney 2005). Here we addressed this issue by examining the pattern of regional phylogenetic signals along the whole-gene alignment of each of our human–macaque quartets. Our analysis indicated that gene conversion seems to be the major cause of homogenization at least in our set of human–macaque quartets.

Materials and Methods

Statistical Tests to Detect Homogenization

The method to detect homogenization is essentially the same as that described in Ezawa et al. (2006). The supplementary file of that paper (http://mbe.oxfordjournals.org/cgi/content/full/msj093/DC1) further elaborates on the detection method. Briefly, our method makes full use of an alignment of quartet sequences in which a paralogous gene pair of two closely related species is used (fig. 1). Because of this choice of quartets, type II informative sites indicating the clustering of intraspecies paralogs (fig. 1*D*) suggest homogenization. We use four statistical tests. 1) The IScomp test examines the abundance of type II sites in the quartet alignment. 2) The T2run test examines whether or not the type II sites segregate from the type I sites, which indicate the clustering of orthologous sequences (fig. 1*C*). 3) The SameTrun test, similar to the T2run test, is also used. 4) The CSrun test exploits the GENECONV program version 1.81 (Sawyer 1989; http://www.math.wustl.edu/~sawyer) to enhance the sensitivity of the whole method. We also use the positive and negative control sets that are generated by computer simulations of quartet evolutions with and without gene conversion, respectively. We integrate the results of computer simulations and the four statistical tests and made a final decision on whether the subject quartet has undergone homogenization or not in a sensitive yet specific way. We would like to call this method, originally proposed by Ezawa et al. (2006), as 4-2-4, for quartets are units of analysis, and two (positive and negative) simulations are conducted, as well as four kinds of statistical tests to detect homogenization.

Both sets of simulations to generate positive and negative control quartets were conducted in almost the same frameworks as those described in the supplementary file of Ezawa et al. (2006), with such parameters as sequence

divergences, base substitution matrices, average codon compositions, and *dN*/*dS* ratios chosen randomly for each quartet according to the distributions observed in the actual sets of quartets. We prepared two sets each of positive and negative control quartets to emulate the actual sets of human–macaque and mouse–rat quartets. For each of the control set, we generated quartet alignments of lengths 450 and 1200 bp. The former and the latter roughly approximate the median lengths of the shorter half and the longer half of the actual quartet alignments, respectively. In order to generate positive control quartets, we simulated gene conversion events whose frequency for each quartet in each species was chosen from Poisson distributions of the mean 1/2, 1, and 3/2 for human–macaque quartets, and those of the mean 1, 2, or 3 for mouse–rat quartets. The length of each gene conversion tract was determined according to the geometric distribution of the mean 200 bp. We excluded a quartet from the positive control set if none of its sequences were changed by gene conversion after speciation.

Dependence of Homogenization Detection Rate on Background Sequence Divergences and Homogenization Tract Length

In order to examine the dependence of homogenization detection rate on quartet properties, we applied our 4-2-4 method and GENECONV version 1.81 (Sawyer 1989) to the sets of positive control quartets that simulate the evolution of human–macaque and mouse–rat quartets under the definite influence of homogenization (see above). We used the SILENT option for GENECONV because the default NUCLEOTIDES option alone shows unexpectedly high false-positive rates when the duplicate sequences undergo different patterns of purifying selection (Ezawa et al. 2006). We took this as a serious problem because the difference in selection patterns between duplicates is expected to be common in evolution after gene duplication (Ohno 1970; Force et al. 1999; Lynch and Force 2000; Lynch and Katju 2004; Katju and Lynch 2006).

After applying the detection methods to positive control sets, we examined how true-positive detection rates depends on quartet properties such as background sequence divergences, which are the divergences that would have resulted if there had been no homogenization after speciation, and the lengths of gene conversion tracts. The divergences were calculated by counting the number of substitutions that actually occurred in the simulation and by dividing it by the sequence length, which always equals the number of nongapped sites because we did not incorporate insertions/deletions into the simulation. We then put quartets into bins each defined by a range of property values, and estimated the detection rate for each of the bins.

Peptide and cDNA Sequences as well as Their Associated Information

We used human and rhesus macaque as the representatives of primates, whereas mouse and rat were used as the representatives of rodents in this study. Figures 2*A* and 2*B* show the phylogenetic relationships of these four species in terms of divergence time and amount of nucleotide substitutions, respectively. It should be noted that the nucleotide substitutions between mouse and rat are almost three times more than between human and macaque. The primates/rodents divergence date and nucleotide divergence values are according to Springer et al. (2003). We referred to Pilbeam (1984), Martin (1993), Takahata and Satta (1997), Glazko and Nei (2003), Steiper et al. (2004), and Steiper and Young (2006) for the time of human–macaque divergence. Regarding the estimated dates of mouse–rat divergence, we consulted Jaegar et al. (1986), Jacobs and Downs (1994), Adkins et al. (2003), and Springer et al. (2003). For the neutral nucleotide divergences, data given in Rat Genome Sequencing Project Consortium (2004), Lindblad-Toh et al. (2005), and Rhesus Macaque Genome Sequencing and Analysis Consortium (2007) were used.

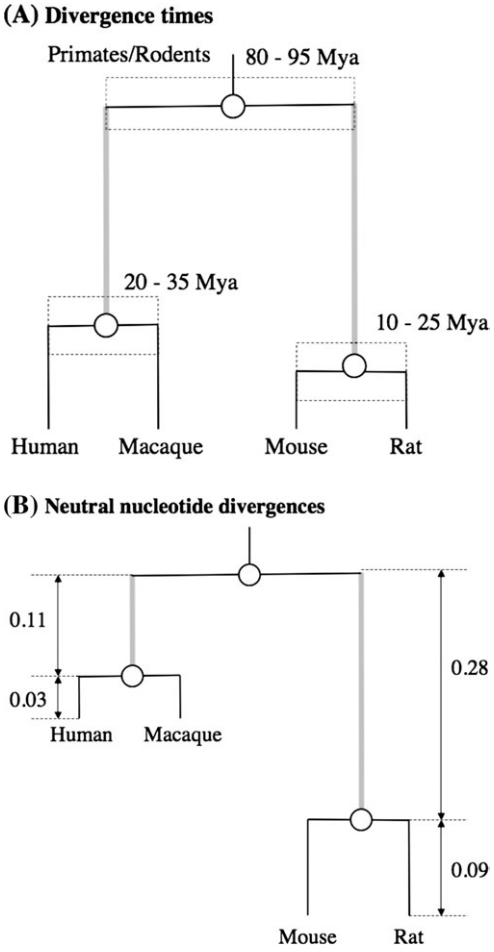

FIG. 2. Evolutionary framework of this study. Shown here is the species tree giving the background of this study. Vertical bars are drawn roughly proportional to the time lengths of the corresponding branches (*A*) or to the corresponding neutral nucleotide divergences (*B*). The quartets used in this study were created by duplication events after the primates–rodents divergence and before the speciation of human–macaque or mouse–rat (thick gray branches). In panel *A*, figures on the right shoulder of each branching point (an open circle) give a range of estimated dates of the divergence event. A dashed box encompassing each branching point displays a time window indicating the range of estimated dates.

We downloaded files of the gene transcript (cDNA) sequences and the peptide sequences predicted on mammalian and avian genomes from the FTP site (ftp.ensembl.org/pub) of the Ensembl database (Hubbard et al. 2007; http://www.ensembl.org) version 43 (updated in February 2007). We obtained data for the following species: human (*Homo sapiens*, 43,738 peptides), rhesus macaque (*Macaca mulatta*, 36,546 peptides), mouse (*Mus musculus*, 32,241 peptides), rat (*Rattus norvegicus*, 33,745 peptides), dog (*Canis familiaris*, 25,568 peptides), cow (*Bos taurus*, 28,334 peptides), opossum (*Monodelphis domestica*, 32,690 peptides), platypus (*Ornithorhynchus anatinus*, 24,763 peptides), and chicken (*Gallus gallus*, 22,186 peptides). Sequences of dog, cow, opossum, platypus, and chicken were used exclusively as outgroups. Mouse and rat sequences were used as outgroups when collecting human–macaque quartets, and human and macaque sequences were used as outgroups when collecting mouse–rat quartets. As for cDNA sequences, we only used those with peptide counterparts, and removed cDNAs of the mitochondrial genes. The genomic map of exons, exon–transcript relationship, transcript–gene relationship, and translation starts and ends of the gene transcripts (cDNAs) were extracted from the MySQL dumps, which in turn were fetched from the above FTP site.

Collecting Human–Macaque Quartets and Mouse–Rat Quartets

Using the cDNA sequences obtained as in the last subsection "Peptide and cDNA Sequences as well as Their Associated Information," we collected genome-wide sets of human–macaque quartets and mouse–rat quartets following a series of procedures largely similar to those described in Ezawa et al. (2006). There are, however, a number of modifications to the technical aspects, such as the software and parameters used for the analyses. Briefly, we first retrieved and screened macaque and rat ortholog candidates of human and mouse cDNAs, respectively, as well as other mammalian or chicken homologs as outgroup sequences. We then searched for and screened human and mouse intraspecies paralogous pairs that are inferred to have duplicated after primates–rodents divergence and before the divergence of human–macaque and mouse–rat, respectively. Finally, we constructed the human–macaque and mouse–rat quartets by combining these intraspecies paralogous pairs and ortholog candidates, and retained only those quartets whose phylogenetic trees indicate the clustering of orthologous pairs. We also removed redundancy due to alternative splicing by only keeping the quartet whose intraspecies paralogous pair exhibits the highest score among the pairs transcribed from each gene pair. These procedures are fully described in sections B to E of the supplementary materials and methods (Supplementary Material online).

After all these procedures, we finally obtained human–macaque quartets and mouse–rat quartets that were generated by duplication events subsequent to the primates–rodents divergence and predating the respective speciation events (fig. 2). The sets of quartets are "nonredundant" in the sense that they never contain two or more cDNA pairs that are produced from the same gene pairs. Moreover, the sets are devoid of evolutionary correlations between quartets due to postspeciation duplication events. Using these two sets of quartets as a basis, we conducted further analyses to detect and characterize homogenization events. For further details, see sections D and E of supplementary materials and methods (Supplementary Material online).

Here we would like to illustrate how our representative quartets are constructed from two species X and Y (X-Y are either human–macaque or mouse–rat for this study). Because our targets are paralogous genes that duplicated before the speciation, a family does not provide a quartet if it consists only of genes duplicated after the speciation of X and Y. Phylogenetic trees in fig. 3*A* and *B* are both consisting of three genes from species X and four genes from species Y. Tree (A) experienced three duplications before the speciation and yields three quartets. All the three quartets should be used for quartet analysis. Gene Y4 in this tree is not involved in quartet because its ortholog, X4, is missing due to either gene loss (genuine evolutionary process) or sequencing/annotation error (artifact). Only one duplication event occurred before speciation and the others occurred after speciation in the phylogenetic tree (B). We can construct eight quartets out of this family shown in tree (B), and the redundant set contains all of them. These eight quartets as well as the homogenization signals on them are, however, historically correlated in an intricate manner due to the postspeciation duplications. Our nonredundant set is devoid of such historical correlation by choosing only one representative (X1, Y2, X3, and Y3), in this case, out of the eight redundant quartets. The representatives are uniquely constructed from the reciprocally best ortholog pairs.

Construction and Masking of cDNA Multiple Alignments

We first constructed peptide alignments via the protein mode of ClustalW version 1.83 (Thompson et al. 1994) with the default setting. We transformed them into the cDNA counterparts replacing amino acids with the corresponding codons, guided by the translational information downloaded from the Ensembl FTP site. We masked dubiously aligned regions to reduce the risk of detecting false signals of homogenization caused by misalignment. We also masked CpG dinucleotides because they are often hypermutable (Ehrlich and Wang 1981) and therefore may cause much more parallel substitutions than expected from naive substitution models. It should be very rare, if ever, that the multiple alignments masked this way should display spurious signs of homogenization. Detailed information on the masking criteria are described in supplementary materials and methods (Supplementary Material online).

Inference of Intraspecies Paralogous cDNA Pairs That Have Undergone Homogenization

Our 4-2-4 method was originally designed to examine whether either of the two intraspecies duplicate pairs in

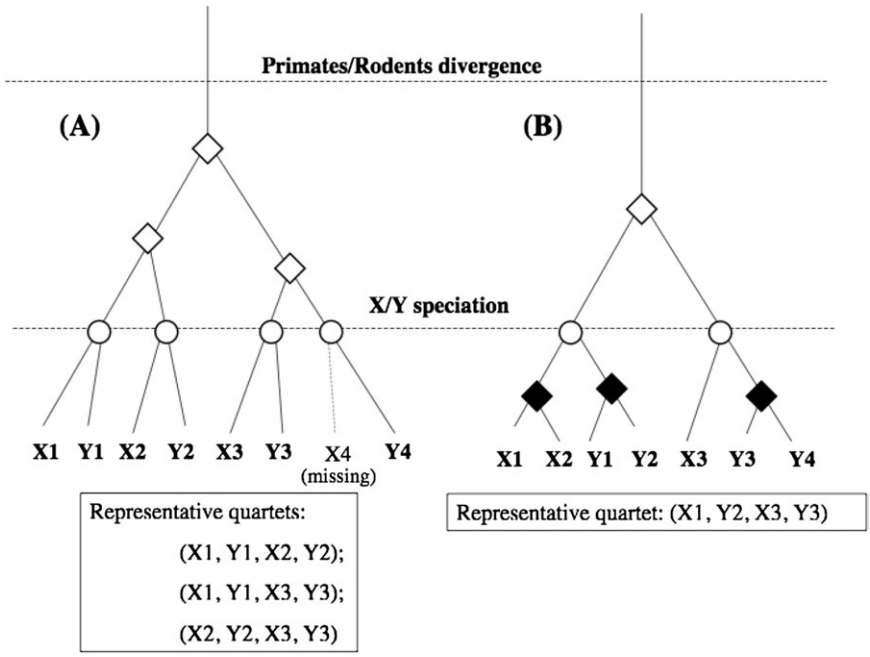

Fig. 3. Illustration of our method to collect representative quartets. To illustrate our method to construct representative quartets, we gave two fictitious gene trees, each consisting of three genes from species X and four genes from species Y. For simplicity, all the duplication events are assumed to have occurred after the primates–rodents divergence (upper dashed horizontal line). Open circles tied with the lower dashed horizontal line represent the speciation of species X and Y. An open diamond and a solid diamond denote a duplication event before speciation and that after speciation, respectively. In tree (*A*), all duplication events occurred before the speciation, and we construct three quartets, all of which are regarded as representatives. We assumed that the ortholog of the gene Y4 in the species X is missing because of gene loss or some other reason. Because of this, the gene Y4 does not contribute to any quartets. Finally in tree (*B*), one duplication event occurred before the speciation but the other duplication events occurred after the speciation. We can construct eight quartets from this family, and all of them are contained in the original redundant set. The nonredundant set, however, contains only one representative quartet out of the eight, in order to avoid the historical correlation of homogenization signals.

a quartet underwent homogenization or not. As it is, our 4-2-4 method cannot distinguish which duplicate pair was homogenized. In order to do so, another series of screening processes are required. The method (Ezawa et al. 2006) basically examines the sequence divergence in the suspected homogenization tract. Briefly, we judged that the human or mouse cDNA pair in a homogenization-positive quartet was homogenized if either 1) GENECONV detected a tract of $P < 0.05$ in that pair, or 2) the best putative gene conversion tract in that pair showed a significantly smaller synonymous distance than the surrounding regions.

Correlations of Homogenization Susceptibility with Properties of Gene Pairs

We classified the quartets according to their linkages, physical distances, and relative transcriptional orientations extracted from the Ensembl MySQL dump, and calculated the prevalence in each category. We defined the physical distance of a quartet as the geometric mean of the distances of the two intraspecies paralogous pair. The distance between a pair of genes is defined as the length of the genomic region between the coding regions of the genes. We divided quartets of various physical distances (L) into seven classes: $L < 50$ kb for class 1, 50 kb $\leq L <$ 100 kb for class 2, 100 kb $\leq L <$ 200 kb for class 3, 200 kb $\leq L <$ 400 kb for class 4, 400 kb $\leq L <$ 800 kb for class 5, and $L \geq 800$ kb for class 6, and unlinked class.

Statistical Tests on the Homogenization Susceptibilities of Functional Categories

Functional categories were assigned to quartets in the refined nonredundant sets based on the results of InterProScan (Zdobnov and Apweiler 2001; http://www.ebi.ac.uk/Tools/InterProScan/) version 4.0. We conducted the following two-step statistical test on each of the human–macaque and mouse–rat sets. We first performed Fisher's exact test to see whether each functional category is significantly more or less prone to homogenization than average in the whole refined set of quartets. For the functional categories with either upper-tailed or lower-tailed $P < 0.05$, we conducted the Cochran–Mantel–Haenszel test (Cochran 1954; Mantel and Haenszel 1959) to control for the effects of sequence divergences and physical proximities (see supplementary materials and methods, Supplementary Material online). Only those categories that also had one-tailed $P < 0.05$ in the Cochran–Mantel–Haenszel test were judged as prone or insusceptible to homogenization.

Construction and Masking of the Multiple Alignment of the Whole-Gene Sequences in a Quartet

We first fetched the chromosomal DNA sequences in the human and mouse genomes from the Ensembl FTP site

(ftp.ensembl.org/pub/). Then, we extracted whole-gene sequences, which include both exons and introns, from the chromosomal sequences with the aid of the transcription start and end coordinates extracted from the MySQL dump of Ensembl database. The whole-gene sequences of each quartet were aligned by the nucleotide mode of ClustalW version 1.83 with parameters dnamatrix = CLUSTALW, gapopen= 8, and gapext = 0.01. In our experiences, this parameter set was known to provide better alignments than default when long gaps are involved, which is often the case with the whole-gene alignment of paralogous sequences. The resulting whole-gene alignment was scanned for regions that appear dubiously aligned, and such sites were masked and neglected in the subsequent analyses. We also masked the CpG dinucleotides in order to reduce false signals of homogenization due to enhanced parallel substitutions. The final product was a masked alignment almost free from spurious signals of homogenization. Details of the masking procedure are provided in supplementary materials and methods (Supplementary Material online).

Tallying the Patterns of Phylogenetic Signals Along the Whole-Gene Alignments

In the whole-gene alignment of each human–macaque quartet, we tried to detect homogenized regions using aggregated type II sites as signals of homogenization. An "aggregated" type II site is defined as a type II site that belongs to a cluster of C or more type II sites: 1) that are not interrupted by type I sites and 2) whose members are not separated from the neighboring member(s) by $L = 200$ sites or more. C is an adjustable parameter that we set 2 or 3 (See section K of supplementary materials and methods, Supplementary Material online, for details). An aggregated type I site is defined similarly by swapping the roles of type I and type II, and it indicates that the region did not undergo homogenization after speciation. It is much less likely that substitutions alone generate aggregated informative sites than isolated informative sites. Aggregated informative sites can therefore be robust signs of genuine phylogenetic relationship or homogenization.

We assigned a pattern of phylogenetic signals as an arrangement pattern of type I and type II tracts along the whole alignment. A type II tract is defined to be a cluster of aggregated type II sites not interrupted by any aggregated type I sites. An aggregated type I tract was defined conversely to a type II tract.

We also examined the distributions of the upper bounds and lower bounds of type II tract lengths. Details on the analyses based on such type I and type II tracts are described in section K of supplementary materials and methods (Supplementary Material online).

Results

True Positive Detection Rates of our 4-2-4 Method and GENECONV

We first examined the rates, or probabilities, that our 4-2-4 method detects homogenization events that actually occurred, and their dependence on the sequence divergences and on the homogenization tract length. For this purpose, we used sets of positive control quartets generated by computer simulations of quartet evolution under the influences of homogenization. Two sets were prepared, one mimicking the set of human–macaque quartets and the other mimicking the set of mouse–rat quartets. For comparison, we also examined the true-positive detection rates of GENECONV (Sawyer 1989) in a parallel manner.

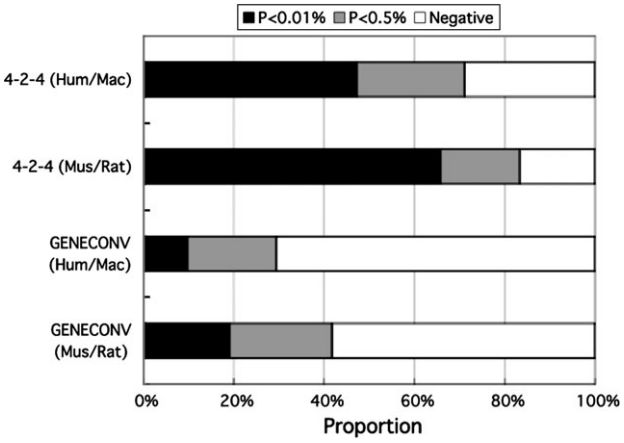

FIG. 4. Overall results of applying our 4-2-4 method and GENECONV to positive control quartets. The four horizontal bar graphs show the results of our 4-2-4 method and GENECONV applied to the 39,413 positive control quartets mimicking actual human–macaque quartets (Hum/Mac) and to the 31,243 positive control quartets mimicking actual mouse-rat quartets (Mus/Rat). In each horizontal bar, black, gray, and white rectangles represent the proportions of extremely positive quartets satisfying $P < 0.01\%$, moderately positive quartets satisfying $P < 0.5\%$, and negative, respectively.

Figure 4 shows the overall results. When applied to the human-macaque-emulating positive controls, our 4-2-4 method showed the true-positive detection rates of 47.3% and 71.1% under the false-positive rates of 0.01% and 0.5%, respectively. GENECONV, on the other hand, showed the true-positive rates of 9.8% and 29.4% under the above false-positive rates. This clearly demonstrates that 4-2-4 substantially outperforms GENECONV regarding sensitivity. This conclusion holds also for the mouse-rat-emulating positive controls, with true-positive detection rates 65.8% and 83.3% for 4-2-4, and 19.1% and 41.8% for GENECONV, under the false-positive rates of 0.01% and 0.5%, respectively (fig. 4).

Another conclusion is that detection rate is higher against the mouse-rat-emulating positive controls than against the human-macaque-emulating ones. This is probably because the former have larger sequence divergences and therefore tend to show stronger signals (such as type II sites) and backgrounds (like type I sites). With our 4-2-4 method, the true-positive detection rate for the mouse–rat data is about 1.2 times that for the human–macaque data. Hereafter in this subsection, the true-positive detection rate is estimated under the false-positive rate of 0.5%.

We then examined the dependence on the average background nucleotide divergence of orthologous pairs (see supplementary fig. S6, Supplementary Material online). Here the "background" means that the divergence is estimated by excluding the effects of homogenization after speciation. As the orthologous divergence increases, the detection rates of our 4-2-4 method gradually increase, whereas that of GENECONV gradually decreases. This is probably because the signal tends to be stronger as the divergence increases, and because the increased rate of recent substitutions tends to disrupt signals (runs of concordant sites) for GENECONV more severely than signals (runs of type II sites) for 4-2-4. It should be noted that our 4-2-4 method maintains fairly high detection rates of >50% across a wide range of orthologous divergence, except the leftmost class of orthologous divergence <0.01 for human–macaque and <0.04 for mouse–rat.

We next examined the dependence on the average background nucleotide divergence of intraspecies paralogous pairs (supplementary fig. S7, Supplementary Material online). Here also our 4-2-4 method consistently performs better than GENECONV throughout the conditions. The detection rate of 4-2-4 gradually increases as the paralogous divergence increase, and it was greater than 50% across a wide range of paralogous divergences, except the divergence <0.08 for human–macaque and <0.12 for mouse–rat.

Dependence on the homogenization tract length was also examined (supplementary fig. S8, Supplementary Material online). As expected, the detection rate generally increases as the tract length increases both for our 4-2-4 method and GENECONV. Our 4-2-4 outperforms GENECONV through the whole range of tract lengths. Especially, 4-2-4 shows relatively high detection rates of around 50% or more against homogenization of tract length between 50 and 200 bp, whereas GENECONV shows poor detection rates against such homogenization.

Generally speaking, our 4-2-4 method outperformed GENECONV under any conditions examined. It showed the true-positive detection rate greater than 50% under a wide range of conditions except with small background orthologous and/or paralogous divergences. This implies that we could conduct relatively reliable correlation analyses after removing the quartets with small sequence divergences. Another important observation is that the detection rate of 4-2-4 shows quite similar dependence patterns on the sequence divergences and on the homogenization tract length when comparing the results against human–macaque and those against mouse–rat (see supplementary figs. S6–S8, Supplementary Material online). This means that observed differences in the values or patterns of homogenization prevalence between primates and rodents detected by 4-2-4, if any, are highly likely to be true biological differences and not due to artifacts such as differences in false-negative rates. This opens the possibility to conduct a fair comparison between primates and rodents regarding susceptibility to homogenization.

Frequency of Homogenization: Overall Average and Dependence on Paralogous Sequence Divergence

We obtained 730 human–macaque quartets and 1,604 mouse–rat quartets that are supposed to have been generated by the duplication after the primates–rodents divergence and before the respective speciation events (thick gray lines in fig. 2). Applying our 4-2-4 method for detecting homogenization to the cDNA alignments of these quartets, we found that 103 human–macaque quartets (14.1%) and 458 mouse–rat quartets (28.6%) are positive for homogenization under the false-positive rate of ca. 0.5%. We were aware, however, that the above set of 730 human–macaque and 1,604 mouse–rat quartets are redundant in the sense that more than one quartet can contain the same duplicate gene, whereas each duplicate pair is contained only once in the sets. This could lead to underestimation or overestimation of the homogenization frequency due to overlaps in the histories of the quartets containing the same genes (evolutionary correlation).

We addressed this issue by constructing a nonredundant set consisting of "representative" quartets that consist of historically independent gene pairs. In this study, our representative quartets consist only of reciprocally best orthologous sequence pairs (see fig. 3 and the related text). This choice has a bonus of retaining only authentic quartets with type I evolutionary history. The resulting nonredundant set consists of 497 quartets for human–macaque and 828 quartets for mouse–rat. Of them, 56 (11.3%) human–macaque quartets and 260 (31.4%) mouse–rat quartets were positive for homogenization. Supplementary tables S3 and S4 (Supplementary Material online) list information on all the quartets in these nonredundant sets of human–macaque and mouse–rat quartets, respectively.

When a human–macaque quartet is positive for homogenization, either the human gene pair or the macaque gene pair or both underwent homogenization. After conducting additional tests to identify gene pairs that underwent homogenization, we estimated that 30 (6.0%) human gene pairs and 179 (21.6%) mouse gene pairs were positive for homogenization. However, these additional tests provide a source of extra false positives/negatives, which makes it obscure how accurate the estimates of positive gene pairs are. Thus, the homogenization prevalence will henceforth be represented with the proportion of positive quartets, which can be estimated more accurately than the proportion of positive gene pairs.

To see the influence of P value thresholds, we divided the category "positive" into two subcategories; extremely positive quartets with $P < 0.01\%$ and moderately positive quartets with $0.01\% < P < 0.5\%$. The numbers of extremely positive quartets in the nonredundant sets were 25 (5.0%) for human–macaque and 137 (16.5%) for mouse–rat. These results indicate that, on average, mouse–rat quartets are more than three times as prone to homogenization as human–macaque quartets, and the ratio is almost uninfluenced by the false-positive rate employed.

We then examined how the homogenization prevalence depends on the average nucleotide divergence between intraspecies paralogs. When we estimated the paralogous divergence of each quartet from the whole coding sequence (CDS) alignment, the prevalence exhibited a strong negative correlation with the paralogous divergence (supplementary fig. S9, Supplementary Material online). This is not surprising, because homogenization by definition reduces the divergence between intraspecies paralogs. In order to see whether less divergent paralogous pairs are more prone to homogenization, we have to examine the dependence on the background paralogous divergence by eliminating the impact of homogenization on the sequence divergence. We approximated such background sequence divergence with the sequence divergence estimated only from the regions whose informative sites are all type I. The correlation did disappear between the homogenization prevalence and the paralogous divergence (supplementary fig. S10, Supplementary Material online). The *P* value of the sequence divergence dependence was 0.99 for human–macaque and 0.29 for mouse–rat under the likelihood ratio test (LRT) with a linear logistic regression model.

However, the strange dependence patterns indicate that some artifacts may be involved. The sudden reduction of the homogenization prevalence at small paralogous divergence (<0.075 for human–macaque, and <0.010 for mouse–rat in supplementary fig. S10, Supplementary Material online) may be due to the poor true-positive detection rate at small paralogous divergence (<0.08 for human–macaque, and <0.12 for mouse–rat; see supplementary fig. S7, Supplementary Material online). Another possible cause would be the "sample exclusion bias," where we retain quartets with small paralogous divergences only when they did not undergo detectable homogenization, because such quartets could erroneously exhibit type II phylogenetic relationships if they experienced homogenization. The anomalous surge of the homogenization prevalence at large paralogous divergences (supplementary fig. S10, Supplementary Material online) may also be due to "sample inclusion bias," where we erroneously include paralogous pairs that duplicated before the primates–rodents divergence, when they underwent intense homogenization recently, because such pairs would show smaller sequence divergences than expected from their actual ages.

In order to mitigate such artificial effects, we further refined our nonredundant set of quartets by discarding quartets with: a) low background orthologous divergences (<0.01 for human–macaque and <0.04 for mouse–rat); b) low background intraspecies paralogous divergences (<0.08 for human–macaque and <0.12 for mouse–rat); or c) high background synonymous distances between intraspecies paralogs (significantly higher than 0.28 for human–macaque and 0.55 for mouse–rat). The conditions (a) and (b) were set according to the analyses on the positive control quartets as described in the previous subsection, and the condition (c) conforms to the phylogenetic screening for duplication events after the primates–rodents divergence (supplementary materials and methods, Supplementary Material online).

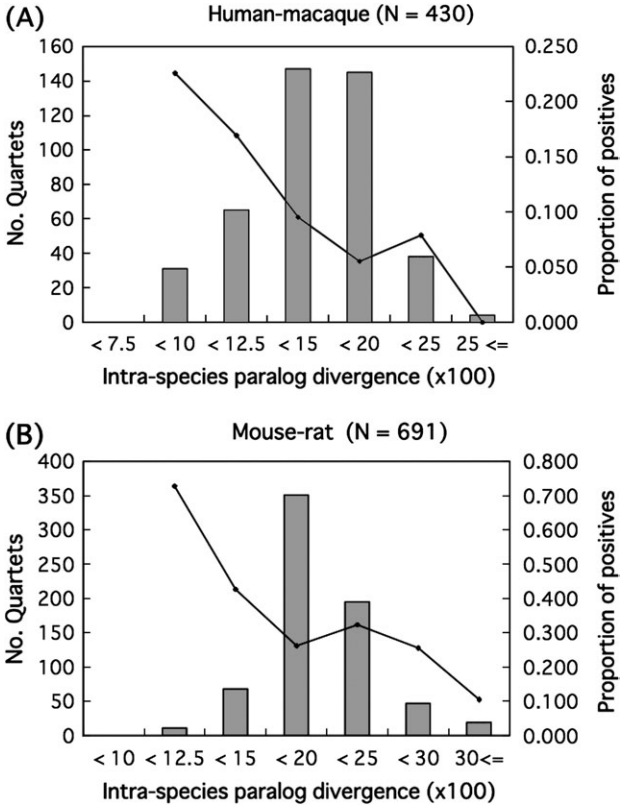

FIG. 5. Dependence of the positive rate on the intraspecies paralogous divergence estimated from the type I regions (after refinements). In each panel, the bar graph shows the numbers of quartets classified by the nucleotide divergence between intraspecies paralogs, and the line graph gives the dependence of the positive rate on the paralogous divergence. The nucleotide divergence was estimated from the regions whose informative sites are all type I, expected to approximate the "background" divergence without the effects of postspeciation homogenization. These are the results after the refinement (see text for details). The panel *A* is for human–macaque quartets, and *B* is for mouse–rat quartets.

The frequencies of positives under these refined sets of quartets at the false-positive rate 0.5% became 10.0% (43 of 430) and 29.8% (206 of 691) for human–macaque and mouse–rat, respectively. Under the false-positive rate of 0.01%, the proportion of positive quartets was 4.0% (17 of 430) for human–macaque and 14.2% (98 of 691) for mouse–rat. These refined quartet sets showed negative correlations between the homogenization prevalence and the background paralogous divergence, as shown in fig. 5; $P = 0.0014$ for human–macaque and 0.017 for mouse–rat under the LRT based on a linear logistic regression model. This is consistent with the past studies that reported that the sequence divergence tends to hamper paralog homogenization (Liskay et al. 1987; Lukacsovich and Waldman 1999; Benovoy and Drouin 2009). Figure 5 also suggests that rodent paralogs are more prone to homogenization than primate paralogs even after controlling for the sequence divergence.

Dependence of Homogenization Prevalence on Physical Proximity between Intraspecies Paralogs

We examined the issue of dependence of homogenization prevalence on physical proximity by using our refined nonredundant sets of quartets and a powerful statistical analysis tool, namely logistic regression analyses. First, we divided each of the two refined sets of quartets into seven categories according to the mean physical proximity between intraspecies paralogs, and estimated the homogenization prevalence in each category, for human–macaque and mouse–rat quartets. Figure 6 and supplementary table S5 (Supplementary Material online) indicate that the prevalence of homogenization positively correlates with physical proximity between intraspecies duplicate copies both for human–macaque and mouse–rat quartets. This is consistent with the result of Ezawa et al. (2006), although their result was devoid of statistical support or consideration of the sequence divergence. In order to take care of these problems, we introduced logistic regression analyses (e.g., Sokal and Rohlf 1995; Agresti 2007). Here, we employ the following regression model for the factor dependence of the logit, $\log(p/(1-p))$, of the positive rate p:

$$[\log(p/(1-p))\sim]\mathrm{Const} + \log(\mathrm{phd}) + \mathrm{Pdiv} + \mathrm{Pdiv}^2 + \log(\mathrm{phd}) \times \mathrm{Pdiv} + \log(\mathrm{phd}) \times \mathrm{Pdiv}^2,$$

where "Const," "phd," and "Pdiv" denote constant, physical distance, and paralogous divergence, respectively, and "×" denotes an interaction. A more detailed description of the model is in section H of supplementary materials and methods (Supplementary Material online). We will henceforth omit the symbol "Const" if there are other terms, because its presence is obvious. In order to examine the significance of the contributions from each term, we conducted an LRT that compares an alternative model with the term in question and a null model without the term. A series of hierarchical LRTs is illustrated in supplementary fig. S12 (Supplementary Material online) using the refined set of human–macaque quartets as an example.

The results of LRTs are summarized in table 1, which shows the statistical significance of the trend represented by each of the terms. It clearly demonstrates that the negative correlation between the homogenization prevalence and the mean physical distance is statistically significant even after controlling for the effects of the paralogous sequence divergence. As for the dependence on the paralogous divergence, the logits of raw nonredundant quartet sets showed a quadratic dependence and the logits of refined sets showed a linear dependence, both of which are weakly significant even after controlling for the effects of physical distance.

The positive correlation between the homogenization prevalence and the physical proximity indicated by fig. 6 and supplementary table S5 (Supplementary Material online) also includes the trend that linked duplicate pairs are more prone to homogenization than unlinked pairs. This trend was also reexamined by taking account of the effects of paralogous nucleotide divergence via the following logistic regression model:

$$\mathrm{Pdiv} + \mathrm{Pdiv}^2 + \mathrm{Lnk} + \mathrm{Lnk} \times \mathrm{Pdiv} + \mathrm{Lnk} \times \mathrm{Pdiv}^2,$$

where "Lnk" represents a linkage status indicator, which equals 0 for a linked pair and 1 for an unlinked pair. The results of hierarchical LRTs under this model show highly significant ($P < 0.1\%$) contributions of the Lnk term (supplementary table S6, Supplementary Material online). This indicates that linked pairs are more prone to homogenization than unlinked pairs even after controlling for the effect of nucleotide divergence, thus confirming the claims made by three studies on mammalian interlocus homogenization (Ezawa et al. 2006; Benovoy and Drouin 2009; McGrath et al. 2009).

Rodent Paralogs are More Prone to Homogenization than Primate ones

The overall prevalence of homogenization for the rodent quartets (29.8%) is about three times higher than that for the primate quartets (10.0%). Visual comparison of the dependence on the paralogous divergence (fig. 5) implies that this is not due to different divergence distributions for these two sets of quartets. Bar graphs in fig. 6 and supplementary table S5 (Supplementary Material online), in contrast, indicate that the set of human–macaque quartets is richer in "physically distant" quartets (distance $\geq$ 800 kb and unlinked) than the mouse–rat set. This may partly explain why the primate quartets appear less prone to homogenization.

When comparing the primate quartets with the rodent ones in the same classes of the physical proximity, however, the latter is still richer in positive quartets, and the difference in the prevalence appears to be significant for four classes (fig. 6 and supplementary table S5, Supplementary Material online). Three classes showing remarkable differences merge together to accommodate linked quartets of physical distances greater than 200 kb, with positive rate less than 10% in human–macaque and more than 20% in mouse-rat. Differences in these categories are fairly significant ($P < 0.5\%$ in Fisher's exact test) and makes the primate and the rodent sets of quartets appear quite different concerning the dependence of the homogenization prevalence on the physical distance. Whereas the mouse–rat quartets can be homogenized considerably even if the distances exceed 800 kb, the human–macaque quartets become relatively immune to gene conversion even with the distance as short as 200 kb (fig. 6). The other significant difference, observed in the class of distance less than 50 kb (with $P = 0.28\%$ for the raw set and 3.0% for the refined set), indicates that even short-range homogenization occurs at a significantly higher frequency in the rodent genome than in the primate genome.

The above observations suggest that such taxonomic difference in the homogenization susceptibility should hold even after controlling for the dependence on the sequence divergence and physical distance. To confirm this

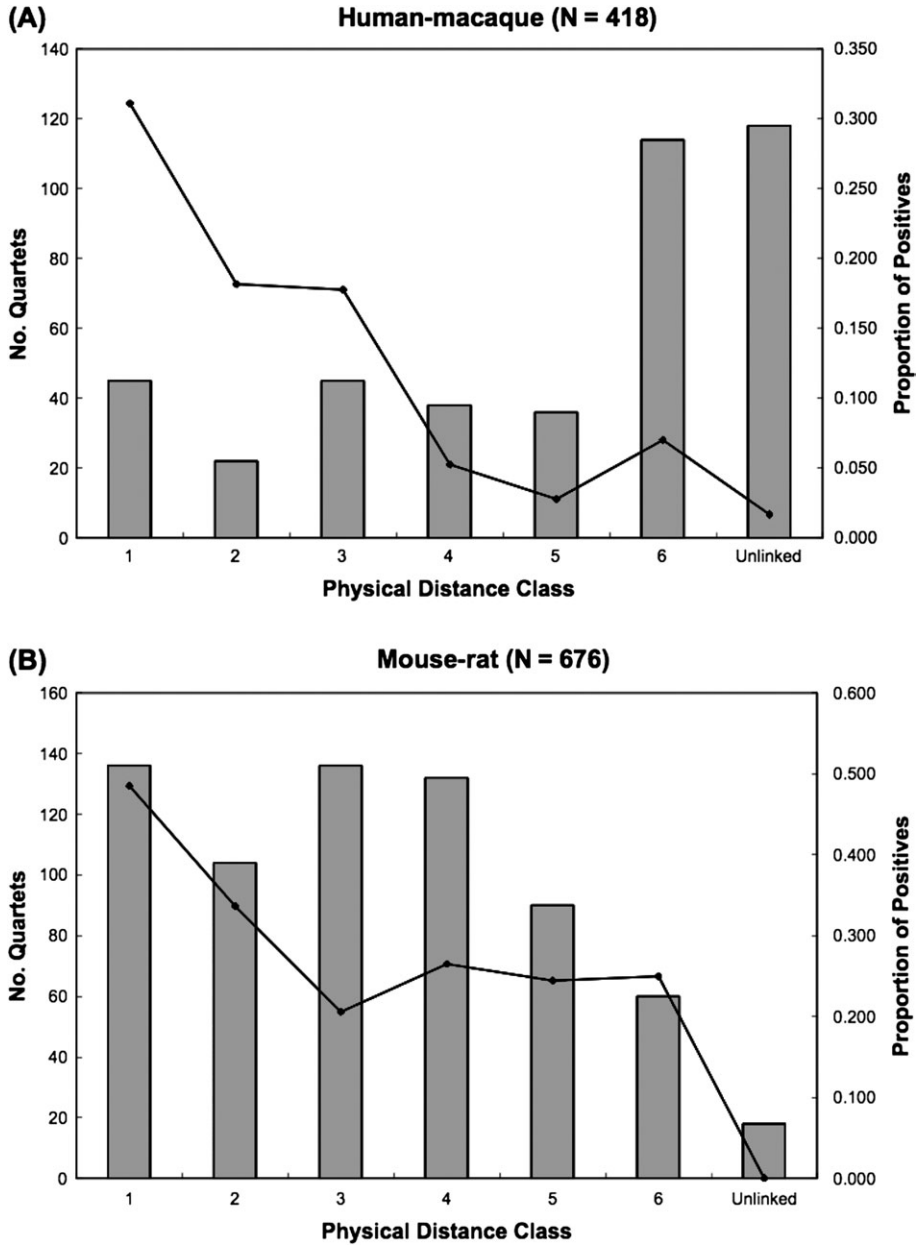

FIG. 6. Positive rates and numbers of quartets in seven categories of physical proximity between intraspecies duplicates (refined sets). In each panel, the bar graph and the line graph show the number of quartets and the proportion of positives, respectively, of each category of physical proximity. We used only those quartets in each of which two intraspecies paralogous pairs have the same linkage status. Panel *A* is for human–macaque and panel *B* is for mouse–rat. "Unlinked" means that two genes in each intraspecies paralogous pair are on different chromosomes. A linked quartet is composed of two intraspecies paralogous pairs each consisting of two genes on the same chromosome. The physical distance of such quartet is defined as the geometric mean of the physical distances of two intraspecies paralogs, each of which in turn is the number of base pairs between the coding sequences of the duplicate genes. The six classes of linked quartets are labeled as follows according to the physical distance (L): class 1 for $L < 50$ kb, class 2 for 50 kb $\leq L <$ 100 kb, class 3 for 100 kb $\leq L <$ 200 kb, class 4 for 200 kb $\leq L <$ 400 kb, class 5 for 400 kb $\leq L <$ 800 kb, and class 6 for 800 kb $\leq$ L. Here we show the results on the refined nonredundant sets of quartets. The results are almost the same even if we used the raw nonredundant sets of quartets (supplementary fig. S11, Supplementary Material online).

expectation, we performed a logistic regression analysis using the model:

$$\log(\text{phd}) + \text{Pdiv} + \text{Pdiv}^2 + \text{Spe} + \text{Spe} \times \log(\text{phd}) + \text{Spe} \times \text{Pdiv} + \text{Spe} \times \text{Pdiv}^2,$$

where "Spe" denotes the species indicator that equals 1 for a human–macaque quartet and 0 for a mouse–rat quartet. The results of the hierarchical LRTs for this model are summarized in table 2, and indicate that the "constant" contributions to the homogenization prevalence differ in a highly significant manner ($P < 1 \times 10^{-7}$) between primates and rodents, whereas the differences in the dependence on the physical distance and on the nucleotide divergence are negligible or marginally significant at best (table 2). Thus, there is definitely a significant difference in the

Table 1. Summary of LRTs with the logistic regression models to examine dependence on the physical distance.

	P values	
Term	**Human–macaque**	**Mouse-rat**
log (phys-dis)	**1.3×10^{-4}*** **	**1.0×10^{-4}*** **
para-div	**0.013***	**0.011***
para-div^2**	**0.40**	**0.71**
log (phys-dis) × para-div	**0.88**	**0.25**
log (phys-dis) × para-div^2**	**0.029**	**0.15**

NOTE.—Results of hierarchical LRTs as shown in supplementary fig. S11 (Supplementary Material online) are summarized. In this analysis, we used only quartets that belong to the refined nonredundant set and both of whose intraspecies paralogous pairs are linked. phys-dis, physical distance between two paralogous genes; para-div, paralogous divergence.

***Significant at 0.02%.

*Significant at 2%.

homogenization frequency in the rodent genome and in the primate genome.

Dependence of Homogenization Prevalence on Relative Orientation

One of the main results of Ezawa et al. (2006) is that the proportion of positive cDNA pairs in mouse does not show a significant correlation with relative transcriptional orientations. Here, we readdressed this issue by taking account of sequence identities via a logistic regression analysis. We only used quartets each having the two intraspecies paralogous pairs with the same relative orientation. We first examined the homogenization prevalence for each relative orientation (table 3). The result indicates that the prevalence does not differ significantly among relative transcriptional orientations, both in the human–macaque and the mouse-rat sets (upper-tailed *P* values: 0.54 for "head-to-tail" vs. the rest, 0.80 for head-to-tail vs. "head-to-head," and 0.31 for head-to-tail vs. "tail-to-tail," in Fisher's exact tests applied to the refined sets of human–macaque quartets; see supplementary table S7, Supplementary Material online for the other sets of quartets and lower-tailed P values).

As a confirmation, we also controlled for the dependence on the paralogous nucleotide divergence and the physical distance using the following logistic regression model:

$$\log(\mathrm{phd}) + \mathrm{Pdiv} + \mathrm{Pdiv}^2 + \mathrm{Ori} + \mathrm{Ori} \times \log(\mathrm{phd}) + \mathrm{Ori} \times \mathrm{Pdiv} + \mathrm{Ori} \times \mathrm{Pdiv}^2,$$

Table 2. Summary of LRTs with the logistic regression models to examine the difference between primates and rodents.

Term	***P* value**
Species	**4.7×10^{-10}**
Species × log (phys-dis)	**0.25**
Species × para-div	**0.33**
Species × para-div^2**	**0.045**

NOTE.—Results of hierarchical LRTs to examine the difference between primates and rodents are summarized. We used only those quartets whose paralogous pairs are both linked. Tests were conducted on a compound set consisting of the refined sets of nonredundant human-macaque and mouse-rat quartets. Species, a species indicator, which equals 1 for primates and 0 for rodents; phys-dis, physical distance between two paralogous genes; para-div, paralogous divergence.

Table 3. Statistics of Quartets with Different Relative Transcriptional Orientations.

Orientation[a]	**Illustration[b]**	**Human–Macaque**	**Mouse–Rat**
Head-to-tail	—> —>	**19/133 (14.3%)[c]**	**134/391 (34.3%)**
Head-to-head	<— —>	**9/50 (18.0%)**	**18/63 (28.6%)**
Tail-to-tail	—> <—	**6/58 (10.3%)**	**24/63 (38.1%)**
Total		**34/241 (14.1%)**	**176/517 (34.0%)**

NOTE.—We used only quartets whose intraspecies paralogous pairs show the same relative orientations.

[a] The key for relative transcriptional orientations: head-to-tail: 5′–3′ 5′–3′, head-to-head: 3′–5′ 5′–3′, tail-to-tail: 5′–3′ 3′–5′.

[b] In this row, a dashed arrow represents the transcriptional orientation of a gene, with the tail and the head of the arrow representing the 5′ and 3′ ends, respectively.

[c] Ineach cell of the first and second columns, the numbers represent: #{positives}/#{quartets} (proportion of positives) in the category considered.

where "Ori" represents an indicator of the relative orientation, which equals 0 for direct (or head-to-tail) pairs and 1 for inverted (or head-to-head or tail-to-tail) pairs. The results of hierarchical LRTs (supplementary table S8, Supplementary Material online) shows that there is no significant contribution involving the relative orientation for any of the quartet sets examined. Therefore, this "lack of dependence on relative orientation" seems to be a genuine biological property that is common to primates and rodents, and maybe across mammals.

Functional Categories Significantly More or Less Prone to Homogenization

Ezawa et al. (2006) found quite a few functional categories that are significantly more or less prone to homogenization than the whole set of mouse cDNA pairs. In this study, we applied the same analysis to the refined nonredundant sets of human–macaque and mouse–rat quartets. After controlling for the effects of the physical proximity and family size (see Materials and Methods), four functional categories, olfactory receptor, Zn-finger B-box, Spla/ryanodine receptor, and guanylate-binding protein, were significantly prone to homogenization for human–macaque, whereas no categories were found to be significantly less prone to homogenization than the whole set (table 4).

The lack of homogenization-insusceptible categories in human may be due to the relatively small number (43) and small proportion (10.0%) of refined nonredundant human–macaque quartets positive for homogenization. Comparing the human–macaque result with that on our new set of mouse–rat quartets (table 4), we found no functional categories that are significantly more or less susceptible to homogenization both in the human–macaque and mouse–rat sets. This may suggest different patterns of selective pressures between primates and rodents. For example, olfactory receptors are more prone to homogenization than average in human–macaque, whereas they do not show significantly deviated homogenization susceptibility in mouse-rat. This may be due to purifying selections against homogenization between rodent olfactory receptors because the olfaction is essential for rodents but not for primates. Family-wise analyses might reveal such differences between primates and rodents in terms of gene functions.

Table 4. Functional Categories that are Significantly More or Less Prone to Homogenization than Average even after Controlling for Sequence Divergence and Physical Proximity.

Domain or Family Name	Total	Positive	*P* value[a]
More prone to homogenization in the human–macaque quartets			
Olfactory receptor (IPR000725)	12	8	0.013
Zn-finger, B-box (IPR000315)	3	3	9.5×10^{-7}
Spla/ryanodine receptor SPRY (IPR003877)	6	4	0.0013
Guanylate-binding protein (IPR003191)	2	2	0.027
Less prone to homogenization in the human–macaque quartets			
None			
More prone to homogenization in the mouse–rat quartets			
Cytochrome P450 (IPR001128)	17	14	7.1×10^{-6}
Cadherin (IPR002126)	5	5	0.0030
Less prone to homogenization in the mouse–rat quartets			
Mammalian taste receptor (IPR007960)	9	0	0.0055
Immunoglobulin V-type (IPR003596)	19	0	0.0062
Vomeronasal receptor type 1 (IPR004072)	62	9	0.017

NOTE.—InterPro IDs assigned to family or domain names are shown in parentheses. Statistics in this table are for the refined nonredundant sets of quartets. Here, we only show functional categories that are significantly more or less prone to homogenization even after controlling for the effects of their sequence divergences and physical proximities.

[a] *P* value in the Cochran–Mantel–Haenszel test (Cochran 1954; Mantel and Haenszel 1959) that controls for the effects of sequence divergences and physical proximities.

Which is Dominant, Unequal Crossing-Over or Gene Conversion?

Although we have examined the prevalence and various properties of homogenization detected in our sets of quartets, so far we have never discussed the mechanisms responsible for it.

Two mechanisms are commonly believed to result in the homogenization of duplicate genes, one is unequal crossing-over (Smith 1976; Ohta 1976) and the other is gene conversion (Jeffreys 1979; Slightom et al. 1980; Ohta 1985; fig. 7). When occurring at an intergenic site, unequal crossing-over just changes the number of duplicate copies (Ritossa and Scala 1969; Schalet 1969). When occurring at an intragenic site, however, it can generate a "chimeric" gene (Donohoue et al. 1989; Lifton et al. 1992; Hampf et al. 2001; Lee et al. 2002; Ezquieta and Luzuriaga 2004), and leads to regionally inconsistent phylogenetic signals that can be detected by our 4-2-4 method (fig. 7*A*). The resulting pattern is expected to be chimeric, consisting of a few relatively long regions containing mostly type I or mostly type II informative sites that are segregated from the other type. A gene conversion event, on the other hand, inherently homogenizes a region or a "tract" of duplicate genes because a tract of a "donor" gene overwrites a homologous region of an "acceptor" gene. This generally results in a "patchy" pattern of regional phylogenetic signals where short tracts of type II informative sites are dispersed in the background of type I informative sites (Kawamura et al. 1992; Kitano and Saitou 1999; Hurles 2001; Winter and Ponting 2005; fig. 7*B*).

Based on this theoretical consideration, we tried to determine the dominant cause of homogenization detected in our 497 nonredundant human–macaque quartets, by examining the patterns of phylogenetic signals the whole-gene alignments display (table 5). In order to reduce false-positive signs of homogenization, the analysis was conducted under two conditions, $C = 2$ and $C = 3$, where C denotes the minimum number of informative sites of the same type in a run that defines a phylogenetic signal (see Materials and Methods for details). Out of the 497 whole-gene alignments, 94 and 45 displayed patterns indicating homogenization under the conditions $C = 2$ and $C = 3$, respectively (table 5). Among the alignments indicating homogenization, only 21 (22%) under $C = 2$ and 15 (33%) under $C = 3$ showed the pattern "12," "21" or "212," which can be explained either as unequal crossing-over or as gene conversion with the same minimum number of events. The remaining 73 (78%) under $C = 2$ and 30 (67%) under $C = 3$, however, require more unequal crossing-over events than gene conversion events if we explain the pattern with the former events alone. It is therefore difficult to explain the statistics obtained here only by means of unequal crossing-over. This indicates that gene conversion should be involved to a considerable degree in the homogenization of most of the quartets showing regional inconsistency of phylogenetic signals.

Actually, the frequency of the cases involving only unequal crossing-over events are further reduced if we note that unequal crossing-over is unlikely to occur between duplicate genes that are unlinked or separated by a long distance. By an analysis detailed in section O of supplementary materials and methods (Supplementary Material online), we estimate that at least 89% ($C = 2$) or 84% ($C = 3$) of homogenized quartets underwent gene conversion.

In order to pursue this issue further, we also obtained distributions of the lengths of type II tracts detected in our 497 human–macaque quartets (table 6; see Materials and Methods). From these distributions, we estimated that at least 86% (119/138) and 72% (46/64) of the tracts detected under the conditions $C = 2$ and $C = 3$, respectively, are estimated to have involved gene conversion in their generation. The detailed arguments are described in section P of supplementary materials and methods (Supplementary Material online).

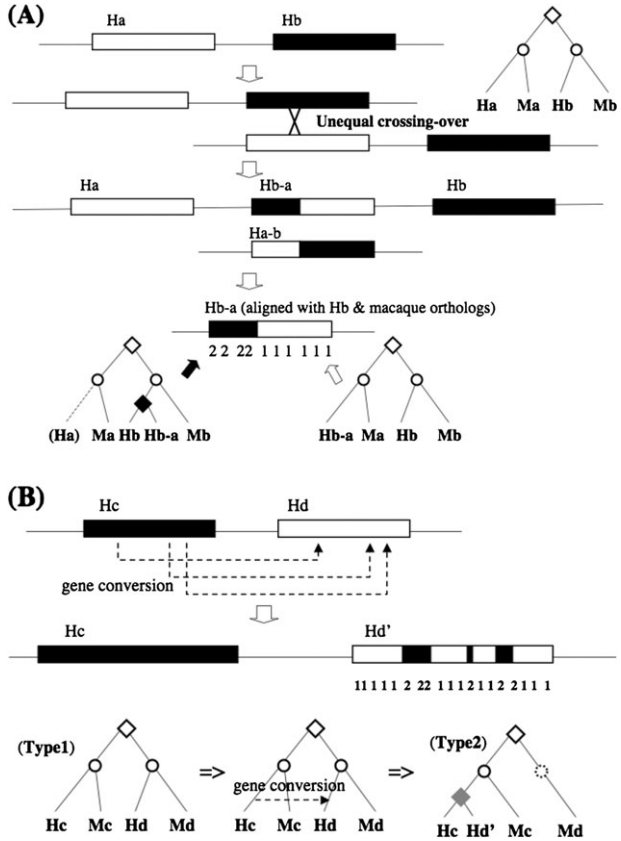

Fig. 7. Schematic illustration of unequal crossing-over, gene conversion, and resulting homogenization patterns. (*A*) Unequal crossing-over at an intragenic site creates a chimeric gene ("Hb-a" in this case). It displays a long-extended homogenization, or a regional inconsistency of phylogenetic signals. (*B*) Gene conversion generates a "patchy" pattern of homogenization. Although for simplicity we only considered one-way gene conversion here, gene conversion can be both ways in actual situations.

Note.—A black solid rectangle and an open (white) rectangle linked by a thin line represent a pair of duplicated genes. A chimeric gene is represented by a rectangle with a mixture of white and black bands. "Hxxx," and "Mxxx" represent a human gene and a macaque gene, respectively. "1" and "2" in these panels represent a type I site and a type II site, respectively. In each phylogenetic tree, an open diamond, an open circle, and a black solid diamond represent a prespeciation duplication event, a speciation event, and a postspeciation duplication event, respectively. A gray solid diamond in panel *B* denotes homogenization caused by gene conversion.

To eliminate the effects of boundaries, we also obtained distributions of type II tract lengths for different patterns of phylogenetic signals (supplementary table S11*A* and *B*, Supplementary Material online), and focused on the pattern 121, which accounts for about a half of the quartets displaying homogenization (table 5). As expected, type II tracts shorter than 500 bp accounted for a majority in the pattern 121 (supplementary table S11*A* and *B*, Supplementary Material online). Judging from these evidences, a majority of homogenization events in our set of human–macaque quartets are likely to be due to gene conversion rather than unequal crossing-over, at least in terms of the number of events.

Table 5. Distributions of the Patterns of Phylogenetic Signals Exhibited Along the Whole-Gene Alignments of the 497 Nonredundant Human–Macaque Quartets.

Phylogenetic Pattern[a]	C = 2[b]	C = 3[b]
None	16	19
1	383	429
2	4	4
Nonhomogenized[c]	403	452
12 or 21	19	15
212	2	0
121	52	20
1212 or 2121	4	3
Five or more blocks[d]	17	7
Homogenized[e]	94	45
Total	497	497

[a] 1, a block of type I sites signaling the normal phylogenetic relationship, namely the clustering of orthologous genes; 2, a block of type II sites signaling homogenization of the duplicate genes from the same species; None, no informative sites providing phylogenetic signals under the condition considered. Details on how to determine the phylogenetic pattern is described in Materials and Methods.

[b] The number of alignments showing each pattern of phylogenetic signals under the specified condition. C, the minimum number of "aggregating" informative sites required for defining a phylogenetic signal. For more details on the conditions, see Materials and Methods.

[c] Subtotal for the patterns NOT indicating homogenization.

[d] The union of quartets showing phylogenetic patterns composed of five or more blocks of type I or type II informative sites.

[e] Subtotal for the patterns indicating homogenization.

Discussion

Comparison of Homogenization between Human–Macaque and Mouse–Rat Quartets

In this study, we collected a nonredundant genome-wide set of 497 human–macaque quartets and another of 828 mouse–rat quartets under comparable conditions and estimated the prevalence of homogenization by applying the 4-2-4 method (Ezawa et al. 2006) to the nonredundant sets of quartets. In the statistical analyses, we did our utmost to mitigate the effects of artifacts by preparing refined sets of quartets on which the false-negative rate is at most below 50%. With such refined nonredundant sets, we detected signs of homogenization between paralogous protein-coding regions in 43 (10.0%) of 430 human–macaque quartets and 206 (29.8%) of 691 mouse–rat quartets. Taking account of the average true-positive rate of 71.1% for human–macaque and 83.3% for mouse–rat, we can estimate the "per-paralog pair" prevalence to be 7.3% $(= 1 - (1 - 0.100/0.711)^{1/2})$ for primates and 19.9% $(= 1 - (1 - 0.298/0.833)^{1/2})$ for rodents, if we assume the independence of homogenization in different species, uniformity of the prevalence, and the equal prevalence in the two closely related species. The actual per-paralog-pair prevalence may be slightly higher because the above assumptions do not usually hold. In the refined nonredundant sets of both primates and rodents, the prevalence of homogenization showed 1) significantly negative correlations with the nucleotide divergence between intraspecies paralogs (fig. 5), 2) significantly positive correlations with the physical proximity (fig. 6) even after controlling for the effects of sequence divergence (table 1), and 3) no significant dependence on the relative transcriptional

Table 6. Distributions of the Lengths of Type II Tracts.

Length (bp)	$C = 2$ (Lower)	$C = 2$ (Upper)	$C = 3$ (Lower)	$C = 3$ (Upper)
1–9	17	0	0	0
10–49	36	7	10	0
50–99	19	4	6	0
100–299	29	24	20	8
300–499	11	19	8	5
500–999	12	17	8	10
1000–1999	5	17	7	15
2000–2999	3	5	1	2
3000–9999	4	15	3	7
10000–	2	30	1	17
Total	138	138	64	64

NOTE.—The lengths of type II tracts in our 497 nonredundant human–macaque quartets were tallied. A "tract" is a merged set of neighboring clusters of aggregated informative sites of the same type. All distributions were obtained under $L = 200$, where L is the upper bound of the distance between informative sites belonging to the same cluster. C is the minimum number of aggregated informative sites that defines a cluster. "Upper" and "Lower" are the upper bound and the lower bound, respectively, of the tract length.

orientation (table 3; supplementary table S7, Supplementary Material online), which holds true even after controlling for the effects of sequence divergence and physical distance between intraspecies paralogs (supplementary table S8, Supplementary Material online).

Although similar analyses were conducted previously regarding the frequency of homogenization in mammalian paralogs (Ezawa et al. 2006; Benovoy and Drouin 2009; McGrath et al. 2009), these analyses did not take account of either the effect of sequence divergence or the effect of false negatives, and therefore the conclusions varied in these previous studies. In this study, we conducted refined statistical analyses, applying logistic regression analyses to the refined sets of quartets. The results of these refined statistical analyses supported the result of Ezawa et al. (2006) and opposed the results of the two recent studies (Benovoy and Drouin 2009; McGrath et al. 2009) concerning the dependence on the physical distance. Regarding the correlation on the paralogous sequence divergence, the analysis supported the conclusion of Benovoy and Drouin (2009), which is also consistent with the past experiments and data analyses (Liskay et al. 1987; Lukacsovich and Waldman 1999). To the best of our knowledge, this study is the first to take account of both sequence divergence and false-negative rate. We are therefore confident that our conclusions are reliable and robust against artifacts, thus serving as a sound basis for future genome-wide analyses of homogenization between duplicates.

We also compared primates and rodents in terms of net homogenization susceptibility. Both lineages shared the positive correlation with the physical proximity as well as no significant correlation with relative orientations. Regarding the overall prevalence, the mouse–rat quartets seem to be about three times more prone to homogenization than the human–macaque quartets (29.8% vs. 10.0% in prevalence). The bar graphs in fig. 6 imply that this disparity is partly explained by the tendency of the primate gene pairs to be farther apart than the rodent pairs. We plan to discuss why the distributions of physical distances between paralogs are different in primates and in rodents (Ezawa K, Ikeo K, Gojobori T, Saitou N, in preparation). Even within the same category of physical proximity, however, the human–macaque quartets still seem less prone to homogenization than mouse–rat quartets (fig. 6), which is significant in four of the seven categories (supplementary table S5, Supplementary Material online). Especially, remarkable disparities were observed in the three subsets containing quartets with paralogous distance >200 kb.

It would be interesting to pursue what caused the huge difference in these distance classes. By a logistic regression analysis, we found that the difference between primates and rodents mainly comes from the constant term and is significant even after controlling for the effects of sequence divergence and physical distance between intraspecies paralogs (table 2). Because the analysis was conducted on the refined sets of nonredundant quartets, the effects of artifacts such as differences in false-negative rates should be negligible, if any.

Theoretically, the prevalence disparity of ca. 3-fold on average and 1.5-fold to 3-fold category-wise between primates and rodents may be reasonable if we consider them as the ratio of the evolutionary distance between mouse and rat to that between human and macaque, as measured by the frequency of gene conversion. The neutral theory of molecular evolution (Kimura 1968, 1983) predicts that the evolutionary distance is roughly proportional to the number of generations separating the two lineages. If we use the genome-wide estimate of the neutral sequence distance as a measure of the evolutionary distance, the distance of 0.174 between mouse and rat (Rat Genome Sequencing Project Consortium 2004) is 2.7 times as large as the distance of 0.065 between human and macaque (Rhesus Macaque Genome Sequencing and Analysis Consortium 2007). Therefore, if the rate of homogenization follows the prediction of the neutral theory, and if other conditions are identical between the two lineages, the prevalence of homogenization in mouse–rat should be about 2.7 times higher than in human–macaque. The observed 3-fold difference in the overall prevalence well matches this theoretical prediction based on the neutral sequence divergence.

We also examined the dependence of homogenization prevalence on functional categories. After removing the effects of physical proximity and sequence divergence, no functional categories are significantly more or less prone to homogenization in both primates and rodents sets of duplicates (table 4). This may indicate different selection patterns on these two lineages.

Signals of Homogenization in the Whole-Gene Alignments of Quartets

In this study, we examined the pattern of phylogenetic signals exhibited by the whole-gene alignment of each quartet. The observed distributions of patterns (table 5), as well as the distributions of type II tract lengths (table 6), seem to be better explained by gene conversion rather than by unequal crossing-over. The proportion of gene conversion

further increased after considering that unequal crossing-over occurs very rarely, if any, between unlinked or physically distant duplicates. This, along with the fairly high prevalence of homogenization (10.0% for primates and 29.8% for rodents), indicates that gene conversion is ubiquitous and can have nonnegligible impacts on the evolution of duplicate genes at least in mammalian genomes.

Results Conflicting with Previous Studies

Some of our results reported in this paper considerably disagree with the conclusions made by Benovoy and Drouin (2009) and McGrath et al. (2009). In the Introduction, we discussed potential problems in their data sets and methods of homogenization detection when examining the dependence of homogenization prevalence on nucleotide divergence or on physical distance. In addition, other aspects merit attention. Benovoy and Drouin (2009) found that the 401 homogenized intrachormosomal pairs they detected were rich in pairs with short distance (typically around or less than 10 kb; fig. 3 of their paper). They found, however, that the correlation was no longer significant after normalizing the number of homogenization events in each class of physical distance by the number of adjacent paralogous gene pairs in the same class. There is an obvious flaw in this analysis: they compared the physical distance distribution for "all" (linked) 401 homogenized paralog pairs with the distribution for only "adjacent" paralog pairs. From their figure 3, we see that the number of such adjacent paralog pairs is only 925, whereas the number of all the linked paralog pairs they examined must have been tens of thousands (but less than the size 55,050 of their whole data set). To conduct a proper correlation analysis, they should either have normalized the physical distance distribution for all linked homogenized paralogs by that for all linked paralogs examined or have normalized the distribution for adjacent homogenized paralogs by that for all adjacent paralogs examined.

We can easily imagine that the distance distribution of all linked paralog pairs would be more broadly spread toward longer distances than the distribution of adjacent paralog pairs, and this is what we actually observed (fig. 6). If they conducted such a proper normalization, Benovoy and Drouin might have observed a significant negative correlation between the homogenization prevalence and the physical distance. Other factors could also hide the real correlation inherent in the data set: a low signal-to-noise ratio and a bad choice of the explanation variable. As discussed in Introduction, the detection rate of homogenization approaches zero when the sequences are almost identical, which causes the low signal-to-noise ratio. Regarding the latter factor, we should note that the logarithmic scale of physical distance enabled us to detect the significant correlation between homogenization prevalence and physical distance. The prevalence is nonzero even at the physical distance exceeding 800 kb (fig. 6). With such a situation, a correlation analysis based on a linear scale of physical distance can easily miss the correlation inherent in the data set. These may be why Benovoy and Drouin (2009) or McGrath et al. (2009) did not detect a significant correlation.

Our Results are Not Likely to be due to Artifacts

Although we found some problems in the previous analyses, this does not necessarily mean that our analyses are free from artifacts or other problems. We first consider our method of constructing nonredundant quartet sets. As explained in fig. 3, the quartet sets we used in our main analyses are nonredundant in the sense that they are devoid of historical correlation after the speciation of the species X and Y. Although some quartets may show overlapping histories before speciation, as in fig. 3*A*, it does not matter because we only analyzed homogenization events after speciation. The problem, if any, would be caused by the case in which, for example, a region of gene X1 homogenizes the homologous regions of genes X2 and X3 in fig. 3*A*. If it happened, the quartet (X2, Y2, X3, Y3) will show a sign of homogenization even if it never underwent homogenization. We conducted a follow-up analysis and identified candidates of such "spurious" homogenization in 5 of 43 positive human–macaque quartets and in 48 of 206 positive mouse–rat quartets (both in the refined nonredundant sets). We reconducted our correlation analyses after removing these quartets showing potentially spurious homogenization, but our main conclusions did not change at all (Supplementary figs. S13 and S14 and supplementary tables S12–S16, Supplementary Material online). This indicates that the effects of artifacts due to our method to collect quartets are negligible, if any.

We also conducted the analysis on the sets of gene families each of which experienced only one duplication event after the primates–rodents divergence. Such families contain at most one quartet per family, and therefore are expected to be completely devoid of spurious homogenization. The homogenization prevalence in these subsets appeared to show no correlation with paralog divergence or with physical distance (supplementary figs. S15 and S16, Supplementary Material online). However, the results were inconclusive because of the small sample sizes (29 for human–macaque and 41 for mouse–rat). If such results hold even when we analyze more such single-quartet families, we will have to examine the causes.

To some readers, our refinement procedure of nonredundant sets of quartets may sound like data manipulation. We are, however, confident that our preparation of refined sets is a scientifically legitimate process. It is common sense in experimental science to use only materials within the range where the detector performs reliably and not to use contaminated materials. Failure to do so would result in erroneous conclusions. Our refinement of the nonredundant quartet sets is just designed to emulate such commonsense experimental practices by weeding out conceivable confounding factors due to the detector's performance characteristics and due to the effects of homogenization. Contrasting the results on the raw data sets (supplementary fig. S10, Supplementary Material online) with those on the refined data sets (fig. 5)

clearly shows that the suspected artifacts actually affected the raw data sets (supplementary fig. S10, Supplementary Material online) but not the refined sets (fig. 5). It should also be noted that, except the dependence on the sequence divergence, our main conclusions remain unchanged regardless of whether the analyses are conducted on the refined sets or on the raw data sets (supplementary tables S17, S18, and S19 and supplementary fig. S11, Supplementary Material online).

In fact, we could have just started from the refined data sets because they are absolutely the data sets that satisfy our two criteria: 1) duplication date should be between the primates–rodents divergence and the speciation (of human–macaque or mouse–rat); 2) the false-negative rate should be less than 50%. It was our deliberate choice that we kept the "strange" patterns on the raw data sets (supplementary fig. S10, Supplementary Material online). By doing so, we intended to show that improperly prepared data sets could result in misleading conclusions.

It is possible that a considerable number of large-scale bioinformatics analyses are actually more or less erroneous either because they are based on improper data sets or because they fail to take account of the detectors' characteristics, such as the false-positive rate and the false-negative rate. We have to be careful not to be deceived by this kind of misleading results.

Although we restricted our analyses to the subsets where the false-negative rate of our 4-2-4 method is less than 50%, there still remains a moderate positive correlation between the true-positive rate and the sequence divergences (supplementary figs. S6 and S7, Supplementary Material online). We do not think, however, this moderate correlation causes serious problems. There are two reasons: 1) the observed homogenization prevalence displays negative correlation with the sequence divergence in spite of the positive correlation between the detection rate and the sequence divergence. This means that the observed negative correlation between the homogenization prevalence and the sequence divergence would be more remarkable if the detection rate were uniform; 2) our logistic regression analyses on the dependence of the prevalence on other factors control for the dependence on the sequence divergence. This should in effect take account of the correlation between the detection rate and the sequence divergence as well.

Putting all these things together, we can conclude that it is highly unlikely that our main results are due to artifacts. This, combined with the problems in the previous analyses (Benovoy and Drouin 2009; McGrath et al. 2009) as discussed in the last subsection, indicates that our results reflect the real biological process rather than artifacts, and therefore are more reliable.

Performance Tests of Homogenization Detectors

One of the main features of this study is that, when refining our nonredundant sets of quartets, we took advantage of the sequence divergence dependence of the true-positive rate of our 4-2-4 method (supplementary fig. S7, Supplementary Material online), which was estimated by applying the method to simulated quartets. It should be noted here that McGrath et al. (2009) also conducted a performance test on GENECONV by applying it to their sets of simulated gene families (table S1 of their paper). Strangely, however, they did not take account of the results of their performance test when conducting the real data analyses. There are also several other differences between their analyses and ours. 1) They used the default NUCLEOTIDES option rather than the SILENT option as we chose. Our choice of the SILENT option is based on our previous finding that the NUCLEOTIDES option of GENECONV could suffer unexpectedly inflated false-positive rates when applied to negative-control quartets whose members underwent different patterns of purifying selection (Ezawa et al. 2006), which is expected to be common in the evolution of duplicate genes (Ohno 1970; Force et al. 1999; Lynch and Force 2000; Lynch and Katju 2004; Katju and Lynch 2006). Our 4-2-4 method, in contrast, displayed stable false-positive rates under any conditions examined. Because the SILENT option examines synonymous changes, which undergoes only weak selection, if any, the use of the SILENT option is expected to mitigate the risk of inflated false-positive rates. 2) We set the threshold P value at 0.5%, whereas McGrath–Casole–Hahn set a more lenient P value threshold of 5%. Our threshold of 0.5% was chosen in order to reduce "contamination" by false positives in the real genome-wide data analyses. If we used the threshold of 5%, there would have been around 25 false positives out of our 497 human–macaque quartets, and around 41 false positives out of our 828 mouse–rat quartets, which could have foiled our correlation analyses. Despite these differences between the two studies, the obtained results broadly agree with each other, especially concerning the poor performance of GENECONV when applied to duplicate pairs with small sequence divergences and/or short homogenization tracts.

Remaining Issues

The refined nonredundant sets of quartets and well-controlled statistical tests employed in this study shed new light on several questions that were left unresolved by our previous study (Ezawa et al. 2006). Yet, there still remain lots of outstanding issues. First, although we showed that primates and rodents share several trends of the prevalence of homogenization, we still do not know how widely these trends are conserved across the phylogenetic tree of life. In order to answer this question, we have to sample pairs of closely related species with sequenced genomes from a wide variety of taxa, and examine how the prevalence of homogenization depends on the properties of gene pairs by the analyses that take account of the effects of false negatives and sequence divergence. This will enable us to compare properties of homogenization in different taxa on a comparable footing, unlike the heterogeneous comparison we tried previously (Ezawa et al. 2006). Because the previous genome-wide studies of gene conversion on a wide variety of species (Semple and Wolfe 1999; Drouin 2002; Gao and Innan 2004; Wang et al. 2007; Xu

et al. 2008) are conducted with different methods of gene pair collection and those of gene conversion detection, mostly without taking good account of the effects of both false negatives and sequence divergences, it would be difficult to sort out real biological differences from artificial effects via a naive comparison of the results of these studies, especially when we try to compare the homogenization prevalence in different species or to conduct sophisticated correlation analyses.

Second, although we defined functional categories based on InterProScan (Zdobnov and Apweiler 2001), family-wise analyses would be more appropriate considering the nontrivial dependence of the prevalence of homogenization on the combination of physical proximity and family size (data not shown). Especially, a sliding-window analysis along the multiple alignment of each gene family would reveal "hot spots" and "cold spots" of homogenization.

Third, to avoid the false-positive/negative problems, we restricted our analyses to the quartets whose phylogenetic trees clearly indicate duplications before the speciation of human and macaque or of mouse and rat. It will be more interesting if we can study quartets whose phylogenetic trees do not reflect their real duplication history any more due to intense homogenization. In an attempt to extend our analyses to such subjects, we applied two-sample runs test (Takahata 1994; supplementary file of Ezawa et al. 2006) to the quartets whose CDS phylogeny is not significantly type I. Fourteen quartets showed significance in the runs test. The CDSs of these 14 quartets are likely to have undergone homogenization that is intensive enough to change the observed phylogeny. They seem to have a wide range of function, from signal receptor through enzymes to transcriptional factors. It would be worth studying them further.

Finally, our ultimate goal is to construct a reasonably accurate model that can describe the evolution of a wide variety of gene families under the influence of homogenization. We need to conduct more sophisticated analyses by integrating evolutionary and comparative genomic approaches as reported here with population genetic approaches (Jeffreys and May 2004; Lindsay et al. 2006; Teshima and Innan 2008) to achieve this goal. Duplicated genes play important roles in evolution of organisms, and their homogenization processes should be analyzed more thoroughly in future studies.

Supplementary Material

Supplementary materials and methods, tables S1–S19, and figures 1–16 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We are grateful to Drs K. Sumiyama and K. Kryukov for discussions on this study. This study was conducted as a part of the Genome Network Project and also as a part of Grant-in-Aid for Scientific Research on Priority Areas "Comparative Genomics" from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

Adkins RM, Walton AH, Honeycutt RL. 2003. Higher-level systems of rodents and divergence time estimates based on two congruent nuclear genes. *Mol Phylogenet Evol*. 26:409–420.

Agresti A. 2007. An introduction to categorical data analysis. 2nd ed. Hoboken (NJ): John Wiley & Sons, Inc.

Arnheim N. 1983. Concerted evolution of multigene families.. In: M Nei, and RK Koehn, editors. Evolution of genes and proteins. Sunderland (MA): Sunauer Associates. p. 38–61.

Balding DJ, Nichols RA, Hunt DM. 1992. Detecting gene conversion: primate visual pigment genes. *Proc R Soc Lond B*. 249:275–280.

Benovoy D, Drouin G. 2009. Ectopic gene conversions in the human genome. *Genomics* 93:27–32.

Brown DD, Wensink PC, Jordan E. 1972. A comparison of the ribosomal DNA's of *Xenopus laevis* and *Xenopus mulleli*: the evolution of tandem genes. *J Mol Biol*. 63:57–73.

Cheung B, Holmes RS, Easteal S, Beacham IR. 1999. Evolution of class I alcohol dehydrogenase genes in catarrhine primates: gene conversion, substitution rates, and gene regulation. *Mol Biol Evol*. 16:23–36.

Cochran WG. 1954. Some methods for strengthening the common χ^2 tests. *Biometrics* 10:417–451.

Donohoue PA, Jospe N, Migeon CJ, Van Dop C. 1989. Two distinct areas of unequal crossingover within the steroid 21-hydroxylase genes produce absence of CYP1B. *Genomics* 5:397–406.

Drouin G. 2002. Characterization of the gene conversions between the multigene family members of the yeast genome. *J Mol Evol*. 55:14–23.

Ehrlich M, Wang RY. 1981. 5-Methylcytosine in eukaryotic DNA. *Science* 212:1350–1357.

Eickbush TH, Eickbush DG. 2007. Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics* 175:477–485.

Ezawa K, OOta S, Saitou N. 2006. Proceedings of the SMBE tri-national young investigator's workshop 2005. Genome-wide search of gene conversions in duplicated genes of mouse and rat. *Mol Biol Evol*. 23:927–940.

Ezquieta B, Luzuriaga C. 2004. Neonatal salt-wasting and 11 b-hydroxylase deficiency in a child carrying a homozygous deletion hybrid CYP11B2 (aldosterone synthase)-CYP11B1 (11 b-hydroxylase). *Clin Genet*. 66:229–235.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–1545.

Gao LZ, Innan H. 2004. Very low gene duplication rate in the yeast genome. *Science* 306:1367–1370.

Glazko GV, Nei M. 2003. Estimation of divergence times for major lineages of primate species. *Mol Biol Evol*. 20:424–434.

Hampf M, Dao NTN, Hoan NT, Bernhardt R. 2001. Unequal crossing-over between aldosterone synthase and 11b-hydroxylase genes causes congenital adrenal hyperplasia. *J Clin Endocr Metab*. 86:4445–4452.

Hubbard TJP, Aken BL, Beal K, et al (45 co-authors). 2007. Ensembl 2007. *Nucl Acids Res*. 35:D610–D617.

Hurles ME. 2001. Gene conversion homogenizes the CMT1A paralogous repeats. *BMC Genomics*. 2:11.

Ibbotson RE, Hunt DM, Bowmaker JK, Mollon JD. 1992. Sequence divergence and copy number of the middle- and long-wave photopigment genes in Old Workd monkeys. *Proc R Soc Lond B*. 247:145–154.

Jackson MS, Oliver K, Loveland J, Humphray S, Dunham I, Rocchi M, Viggiano L, Park JP, Hurles ME, Santibanez-Koref M. 2005.

Evidence for widespread reticulate evolution within human duplicons. *Am J Hum Genet*. 77:824–840.
Jacobs LL, Downs WR. 1994. The evolution of murine rodents in Asia. In: Tomida Y, Li C, Setoguchi T, editors. Rodents and lagomorph families of Asian origin and diversification. Tokyo (Japan): National Science Musium Monograph. p. 149–156.
Jaegar JJ, Tong H, Denys C. 1986. The age of Mus-Rattus divergence: paleontological data compared with the molecular clock. *C R Acad Sci Paris*. 302(ser. II):917–922.
Jeffreys A. 1979. DNA sequence variants in the G gamma-, A gamma-, delta-, and beta- globin genes of man. *Cell* 18:1–10.
Jeffreys AJ, May CA. 2004. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat Genet*. 36:151–156.
Katju V, Lynch M. 2006. On the formation of novel genes by duplication in the *Caenorhabditis elegans* genome. *Mol Biol Evol*. 23:1056–1067.
Kawamura S, Saitou N, Ueda S. 1992. Concerted evolution of the primate immunoglobulin a-gene through gene conversion. *J Biol Chem*. 267:7359–7367.
Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* 217:624–626.
Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.
Kitano T, Saitou N. 1999. Evolution of Rh blood group genes have experienced gene conversions and positive selection. *J Mol Evol*. 16:111–120.
Lee H-H, Niu D-M, Lin R-W, Chan P, Lin C-Y. 2002. Structural analysis of the chimeric CYP1P/CYP21 gene in steroid 21-hydroxylase deficiency. *J Hum Genet*. 47:517–522.
Li W-H. 1997. Molecular evolution. Sunderland (MA): Sinauer Associates.
Lifton RP, Dluhy RG, Powers M, Rich GM, Cook S, Ulick S, Lalouel JM. 1992. A chimaeric 11b-hydroxylase/aldosterone synthase gene causes glococorticoid-remediable aldosteronism and human hypertension. *Nature* 355:262–265.
Lindblad-Toh K, Wade CM, Mikkelsen TS, et al (46 co-authors) 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438:803–819.
Lindsay SJ, Khajavi M, Lupski JR, Hurles ME. 2006. A chromosomal rearrangement hotspot can be identified from population genetic variation and is coincident with a hotspot for allelic recombination. *Am J Hum Genet*. 79:890–902.
Liskay RM, Letsou A, Stachelek JL. 1987. Homology requirement for efficient gene conversion between duplicated chromosomal sequences in mammalian cells. *Genetics* 115:161–167.
Lukacsovich T, Waldman AS. 1999. Suppression of intrachromosomal gene conversion in mammalian cells by small degrees of sequence divergence. *Genetics* 151:1559–1568.
Lynch M, Force A. 2000. The probability of duplicated gene preservation by subfunctionalization. *Genetics* 154:459–473.
Lynch M, Katju V. 2004. The altered evolutionary trajectories of gene duplicates. *Trends Genet*. 20:544–549.
Mantel N, Haenszel W. 1959. Statistical aspects of analysis of data from retrospective studies of disease. *J Natl Cancer Inst*. 22:719–748.
Martin RD. 1993. Primate origins: plugging the gaps. *Nature* 363:223–234.
McGrath CL, Casola C, Hahn MW. 2009. Minimal effect of ectopic gene conversion among recent duplicates in four mammalian genomes. *Genetics* 182:615–622.
Nathans J, Piantanida TP, Eddy RL, Shows TB, Hogness DS. 1986. Molecular genetics of inherited variation in human color vision. *Science* 232:203–210.
Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet*. 39:121–152.
Ohno S. 1970. Evolution by gene duplication. Berlin (Germany): Springer-Verlag.
Ohta T. 1976. Simple model for treating evolution of multigene families. *Nature* 262:74–76.
Ohta T. 1985. A model of duplicative transposition and gene conversion for repetitive DNA families. *Genetics*. 110:513–524.
Petes TD, Hill CW. 1988. Recombination between repeated genes in microorganisms. *Annu Rev Genet*. 22:147–168.
Pilbeam DR. 1984. The descent of hominids and hominoids. *Sci Am*. 250:60–69.
Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the brown Norway rat yields insights into mammalian evolution. *Nature* 428:493–521.
Rhesus Macaque Genome Sequencing and Analysis Consortium. 2007. Evolutionary and Biological Insights from the Rhesus Macaque Genome. *Science* 316:222–234.
Ritossa FM, Scala G. 1969. Equilibrium variations in the redundancy of rDNA in *Drosophila melanogaster*. *Genetics* 61(Suppl):305–317.
Sawyer S. 1989. Statistical tests for detecting gene conversion. *Mol Biol Evol*. 6:526–538.
Schalet A. 1969. Exchanges at the bobbed locus of *Drosophila melanogaster*. *Genetics*. 63:133–153.
Schienman JE, Holt RA, Auerbach MR, Stewart CB. 2006. Duplication and divergence of 2 distinct pancreatic ribonuclease genes in leaf-eating African and Asian colobine monkeys. *Mol Biol Evol*. 23:1465–1479.
Scott AF, Heath P, Trusko S, Boyer SH, Prass W, Goodman M, Czelusniak J, Chang L-Y.E, Slightom JL. 1984. The sequence of the gorilla fetal globin genes: evidence for multiple gene conversions in human evolution. *Mol Biol Evol*. 1:371–389.
Semple C, Wolfe KH. 1999. Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. *J Mol Evol*. 48:566–576.
Shyue S-K, Li L, Chang BH-J, Li W-H. 1994. Intronic gene conversion in the evolution of human X-linked color vision genes. *Mol Biol Evol*. 11:548–551.
Slightom JL, Blechl AE, Smithies O. 1980. Human fetal G gama- and A gamma-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. *Cell* 21:627–638.
Slightom JL, Chang LY, Koop BF, Goodman M. 1985. Chimpanzee fetal G gamma- and A gamma-globin gene nucleotide sequences provide further evidence of gene conversions in hominine evolution. *Mol Biol Evol*. 2:370–389.
Smith GP. 1976. Evolution of repeated DNA sequences by unequal crossover. *Science* 191:528–535.
Sokal RR, Rohlf FJ. 1995. Biometry, 3rd ed. New York: W.H. Freeman and Company.
Springer MS, Murphy WJ, Eizirik E, O'Brien SJ. 2003. Placental mammal diversification and the Crataceous-Tertiary boundary. *Proc Natl Acad Sci U S A*. 100:1056–1061.
Steiper ME, Young NM. 2006. Primate molecular divergence dates. *Mol Phylogenet Evol*. 41:384–394.
Steiper ME, Young NM, Sukarna TY. 2004. Genomic data support the hominoid slowdown and an early Oligocene estimate for the hominoid-cercopithecoid divergence. *Proc Natl Acad Sci U S A*. 101:17021–17026.
Takahata N. 1994. Comments on the detection of reciprocal recombination or gene conversion. *Immunogenetics* 39:146–149.
Takahata N, Satta Y. 1997. Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. *Proc Natl Acad Sci U S A*. 94:4811–4815.
Teshima KM, Innan H. 2008. Neofunctionalization of duplicated genes under the pressure of gene conversion. *Genetics* 178:1385–1398.
Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment

through sequence weighting, positions-specific gap-penalties and weight matrix choice. *Nucl Acid Res*. 22:4673–4680.

Wang X, Tang H, Bowers JE, Feltus FA, Paterson AH. 2007. Extensive concerted evolution of rice paralogs and the road to regaining independence. *Genetics* 177:1753–1763.

Winter EE, Ponting CP. 2005. Mammalian BEX, WEX and GASP genes: coding and non-coding chimaerism sustained by gene conversion events. *BMC Evol Biol*. 5:54.

Xu S, Clark T, Zheng H, Vang S, Li R, Wong GK-S, Wang J, Zheng X. 2008. Gene conversion in the rice genome. *BMC Genomics*. 9:93.

Zdobnov EH, Apweiler R. 2001. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17:847–848.

Zhou Y-H, Li W-H. 1996. Gene conversion and natural selection in the evolution of X-linked color vision genes in higher primates. *Mol Biol Evol*. 13:780–783.