Proceedings of the SMBE Tri-National Young Investigators' Workshop 2005

Genome-Wide Search of Gene Conversions in Duplicated Genes of Mouse and Rat

Kiyoshi Ezawa, Satoshi OOta,[1] *and Naruya Saitou*

Division of Population Genetics, National Institute of Genetics, Mishima, Japan

Gene conversion is considered to play important roles in the formation of genomic makeup such as homogenization of multigene families and diversification of alleles. We devised two statistical tests on quartets for detecting gene conversion events. Each "quartet" consists of two pairs of orthologous sequences supposed to have been generated by a duplication event and a subsequent speciation of two closely related species. As example data, EnsEMBL mouse and rat cDNA sequences were used to obtain a genome-wide picture of gene conversion events. We extensively sampled 2,641 quartets that appear to have resulted from duplications after the divergence of primates and rodents and before mouse-rat speciation. Combination of our new tests with Sawyer's and Takahata's tests enhanced the detection sensitivity while keeping false positives as few as possible. About 18% (488 quartets) were shown to be highly positive for gene conversion using this combined test. Out of them, 340 (13% of the total) showed signs of gene conversion in mouse sequence pairs. Those gene conversion–positive gene pairs are mostly linked in the same chromosomes, with the proportion of positive pairs in the linked and unlinked categories being 15% and 1%, respectively. Statistical analyses showed that (1) the susceptibility to gene conversion correlates negatively with the physical distance, especially the frequency of 29% was observed for gene pairs whose distances are smaller than 55 kb; (2) the occurrence of gene conversions does not depend on the transcriptional direction; (3) small gene families consisting of between three and six contiguous genes are highly prone to gene conversion; and (4) frequency of gene conversions greatly varies depending on functional categories, and cadherins favor gene conversion, while vomeronasal receptors type 1 and immunoglobulin V–type proteins disfavor it. These findings will be useful to deepen the understanding of the roles of gene conversion.

Introduction

Gene conversion, also called nonreciprocal recombination, is a process where a tract of DNA overwrites a homologous one (e.g., Petes and Hill 1988; Haber 2000). According to the positional relationship of the pair of tracts, gene conversion is classified as (1) intrachromatid, (2) sister chromatid, (3) classical (allelic), (4) semiclassical (nonallelic between homologous chromosomes), and (5) ectopic (heterochromosomal) (Li 1997, pp. 310–311). The classical conversion is the interaction between alleles of the same locus (allelic gene conversion), whereas the others involve two different loci (interlocus gene conversion).

There are many studies discussing the evolutionary implications of gene conversion, and they are broadly classified into two groups. One is the enhancement of allelic diversity via gene conversion, either allelic or interlocus. This has been documented for genes involved in immune response, like major histocompatibility complex genes (e.g., Weiss et al. 1983; Kuhner et al. 1991; Martinsohn et al. 1999; Richman et al. 2003; Reusch, Schaschi, and Wegner 2004), and genes controlling self-incompatibility (Charlesworth et al. 2003). The other is the homogenization or concerted evolution of multiple-gene families via interlocus gene conversion, possibly in cooperation with unequal crossing-over. The examples are the rDNA multigene family (Arnheim et al. 1980; Arnheim 1983), red and green opsin genes in Old World monkeys (e.g., Ibbotson et al. 1992; Winderickx et al. 1993; Shyue et al. 1994; Zhou and Li 1996), and genes controlling self-incompatibility (e.g., Cabrillac et al. 1999; Prigoda, Nassuth, and Mable 2005). The homogenization of gene families can be a severe obstacle to the currently dominant paradigm that gene repertoire expands through neofunctionalization and subfunctionalization of duplicated genes (Ohno 1970; Force et al. 1999; Lynch and Force 2000). Therefore, knowing the genomic prevalence and frequency of gene conversion is indispensable for estimating the potential or speed of the genome evolution.

Detection of gene conversion is also important for the molecular evolutionary study in general. Many instances of gene conversion were revealed by the regional inconsistencies of gene phylogenies (e.g., Scott et al. 1984; Kawamura, Saitou, and Ueda 1992; Cheung et al. 1999; Kitano and Saitou 1999). Therefore, we may incorrectly infer the duplication dates and, consequently, the phylogenetic relationships between paralogous genes. This can happen whenever the effects of gene conversion overpower phylogenetic signals, as exemplified by some instances of the concerted evolution mentioned above.

In spite of many instances of gene conversion documented so far, their anecdotal nature seems to let some researchers believe that gene conversion is a rarity, and therefore, a naive phylogenetic picture still holds true for most of the duplicated genes. So far, genome-wide searches for gene conversion have been performed only for nematode worm (Semple and Wolfe 1999) and yeast (Drouin 2002; Gao and Innan 2004). For vertebrates, the prevalence of gene conversion has been estimated so far only in a small-scale analysis before the advent of the genome era (Shields 2000). Now that we have a number of mammalian genome sequences available (Mouse Genome Sequencing Consortium

[1] Present address: Department of Biological Systems, Bioresource Center, RIKEN Tsukuba Institute, Tsukuba, Japan.

Key words: gene conversion, duplicated genes, mouse, rat, genome-wide analysis.

E-mail: nsaitou@genes.nig.ac.jp.

Mol. Biol. Evol. 23(5):927–940. 2006
doi:10.1093/molbev/msj093
Advance Access publication January 11, 2006

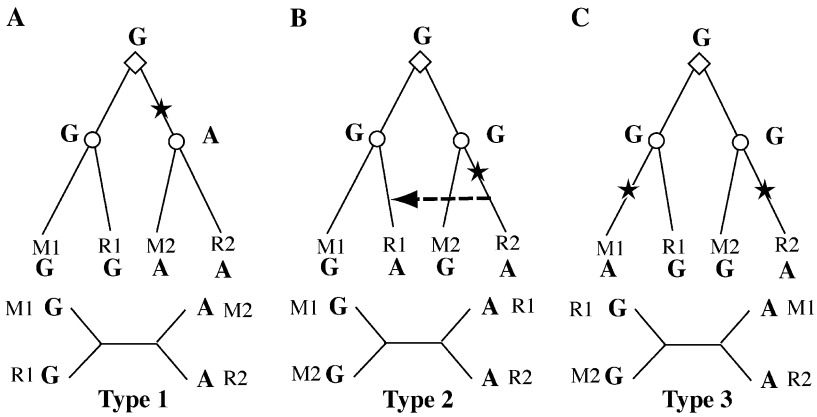

FIG. 1.—A quartet and its informative sites (*A*). Top: A duplication event (an open diamond) and the subsequent divergence of two species (open circles), mouse and rat in our case, create a quartet of sequences mouse1, rat1, mouse2, and rat2, which are abbreviated as M1, R1, M2, and R2, respectively. The commonest informative sites are mainly produced by a base substitution (a star), from a guanine (G) to an adenine (A) in this example, on an internal branch. Bottom: This type of informative sites, called the type 1 sites here, indicates the clustering of orthologous sequences. (*B*) A base substitution followed by a gene conversion event (a broken arrow) can create informative sites that indicate the clustering of two sequences in the same genome. We will call them the type 2 sites here. (*C*) A third kind of informative sites, the type 3 sites, is generated by parallel substitutions.

2002; International Human Genome Sequencing Consortium 2004; Rat Genome Sequencing Project Consortium 2004), we can use these resources to obtain a genome-wide view of gene conversion events, especially regarding how common they are, and what kind of gene pairs are prone to gene conversion. The results of such analyses should become precious assets for studies on the evolution of duplicated genes.

Here, led by the above motivations, we searched the whole sets of mouse and rat cDNA sequences in the EnsEMBL database (Hubbard et al. 2002, 2005; Birney et al. 2004; http://www.ensembl.org/index.html) for traces of gene conversion. Our gene conversion detection is based on two statistical tests conducted on quartets. A "quartet" consists of two pairs of orthologous sequences supposed to have been generated by a duplication event followed by the speciation of two closely related species, mouse and rat in our case (see fig. 1). The genome-wide divergence between mouse and rat is approximately 17% (Abe et al. 2004; Rat Genome Sequencing Project Consortium 2004), and this intermediate species divergence is expected to be suitable for the quartet-based methods. Although the combination of these tests gives a fairly strict screening with very low false-positive rate, some types of gene conversion events turned out to escape the detection by our method. So we combined our statistical tests with Sawyer's (1989) test as well as with Takahata's (1993) in a manner that enhances the detection sensitivity and yet keeps the low rate of false positives. The Supplementary File details this theoretical part (Supplementary Material online).

The frequency of false positives among gene conversion candidates was not considered in the past large-scale gene conversion searches. However, we think it crucial to always estimate the false-positive rate. We addressed this issue by simulating the quartet evolution in two different ways, one with gene conversion and the other without. These simulation results also guided us to the optimum combination of the four statistical tests mentioned above (see Supplementary File, Supplementary Material online).

We also examined correlations between the prevalence of gene conversion and physical and biological properties of gene pairs, in order to figure out what kind of gene pairs are prone to gene conversion.

Materials and Methods

Peptide and cDNA Sequences as well as Their Associated Information

We retrieved files of the gene transcript (cDNA) sequences and the peptide sequences predicted on mammalian and avian genomes from the file transfer protocol (FTP) site (ftp://ftp.ensembl.org/pub) of the EnsEMBL database (Hubbard et al. 2002, 2005; Birney et al. 2004; http://www.ensembl.org/index.html) version 27 (updated in December 2004). We obtained data for the following species: mouse (*Mus musculus*, 32,442 peptides), rat (*Rattus norvegicus*, 28,545 peptides), human (*Homo sapiens*, 34,111 peptides), dog (*Canis familiaris*, 30,308 peptides), and chicken (*Gallus gallus*, 28,416 peptides). Sequences of the latter three species are used as outgroup. As for the cDNA sequences, we only used those with peptide counterparts. We also fetched mysql dumps of information on the gene transcripts from the FTP site above, and then we extracted information on the genomic map of exons, exon-transcript relationship, transcript-gene relationship, and translation starts and ends of the gene transcripts (cDNAs).

Rat Ortholog Candidates and Outgroup Sequences of Mouse cDNAs

We conducted BlastP searches (Altschul et al. 1990) using mouse peptides as queries and a target database consisting of peptides of other mammals with the BLOSUM62 scoring matrix and the E value threshold of 1×10^{-5}. Then we chose subject sequences whose homologous segment pairs (HSPs) showed more than 35% sequence identity and were 100 amino acids or longer. We did this first screening in order to secure the reliability of the subsequent sequence analyses by restricting our subjects to well above the upper

bound of the "twilight zone," below which sequence alignments are often unreliable (Doolittle 1986; Rost 1999). The *E* value threshold of 1×10^{-5} is large enough to retrieve most of the sequences satisfying the above conditions. Out of this set of homolog candidates, we chose the rat ortholog candidate of each mouse cDNA in the following manner.

We first picked out rat peptide sequences that scored the best. However, because our main targets are members of multigene families, the best-scoring sequence can sometimes be different from the real ortholog. This can occur when the orthologous sequences have undergone disparate functional constraints in different species. We therefore kept those sequences that are "close to" the best sequence and left the judgment of orthology to the final screening using nucleotide alignments. We regarded the sequences as close to the best sequence if they satisfy the inequality on the effective distance, *d_eff*: $d\_eff\ (sb) <= d\_eff\ (best) + 2 * \{d\_eff\ (best)\}^{1/2}$, where "*sb*" and "*best*" stand for the subject and the best-scoring sequence, respectively, and we defined *d_eff* (*sb*) by –log (*score*(*qu*,*sb*)/*score*(*qu*,*qu*)). Here *score*(*a*,*b*) is the score of the best HSP between the sequences *a* and *b*, and "*qu*" stands for the query sequence. The point here is to take account of some fluctuation in sequence divergences. And we expect this effective distance should be a good proxy of the sequence divergence because it accounts for substitutions, insertions/deletions, and the alignment coverage.

We then constructed pairwise nucleotide alignments of the cDNA counterparts of the query and subject sequences in a manner similar to the construction of multiple alignments described below. We chose the rat cDNA sequence whose nucleotide alignment with the query sequence scores the best among the surviving subjects. Finally, the best-scoring subject was kept as the "ortholog candidate" if the following two criteria are met: (1) the unambiguously aligned regions excluding gaps cover 70% or more of the maximum between the query and the subject lengths and (2) the number of synonymous differences per synonymous site is 0.3 or less. We counted synonymous sites and synonymous differences by the "dists" program of the ODEN package (Ina 1994), which implements the method of Nei and Gojobori (1986). In this way, we found rat ortholog candidates for 20,910 mouse peptide-coding cDNAs.

We also applied similar procedures to human and dog sequences with a more lenient threshold of 0.6 for synonymous differences per synonymous site, which yielded human and dog outgroup sequences for 20,278 and 19,405 of mouse cDNAs, respectively.

Intraspecies Paralogous Sequences of Mouse cDNAs

The main purpose here is to retrieve, from the mouse genome, pairs of paralogous sequences that have duplicated after the rodent-primate divergence. To speed up the filtering process, we introduced an appropriate "outgroup" species that defines a "natural cutoff." In contrast to numerical cutoffs, a natural cutoff is given by the best score between the query sequence and the sequences from the outgroup species. Because the functional constraints on amino acid substitutions vary depending on the proteins, the natural cutoff should be more suitable to retrieve sequences that diverged from the query after a specified taxonomical event. We chose chicken as the outgroup most suitable for the purpose here. Humans and dogs are too close considering that the mouse genome evolved two to three times faster than the human genome, in terms of branch lengths (Rat Genome Sequencing Project Consortium 2004). Birds and mammals diverged about 310 MYA (Hedges 2002; Reisz and Muller 2004), about three times older than the primate-rodent divergence (Springer et al. 2003). It should therefore be very rare that this natural cutoff rejects our target here.

We started with BlastP searches using mouse peptides as queries and a target database made of mouse and chicken peptides with the parameters used above. Similar to the above section, we first prepared a set of homologs using the thresholds of 35% identity and 100 amino acids. Then we screened mouse paralogous sequences for each query according to the following procedures.

First, we selected only mouse peptides whose alignments cover 70% or more of the maximum between the query and subject lengths. Then, for each mouse query sequence, we used the best score among the chicken homologs as a natural cutoff and discarded mouse sequences scoring lower than that. At this stage, we unconditionally accepted mouse pairs without any chicken homologs. This screening yielded 112,711 mouse pairs.

The next step is to screen the surviving mouse pairs by using the phylogenetic relationship with human and dog outgroup sequences. For each pair of mouse cDNAs, after adding its best human and dog homologs as outgroup sequences, we constructed a multiple alignment via the method described below. Then we constructed a phylogenetic tree using the neighbor-joining method (Saitou and Nei 1987) and a distance matrix estimated by Kimura's (1980) two-parameter model that are implemented in ClustalW (Thompson, Higgins, and Gibson 1994) version 1.83. Finally, we selected only those mouse pairs that satisfy the following criteria: (1) synonymous differences are less than 0.6 per synonymous site, (2) the unrooted tree topology shows clustering of two mouse genes if they have both human and dog homologs, and (3) the synonymous distance between the mouse sequences is less than 1.5 times the minimum synonymous distance between the mouse and the outgroup sequences. Here the synonymous distances are measured with dists of the ODEN package.

After filtering out the pairs each consisting of alternative splicing variants of the same gene, we obtained 42,546 pairs of mouse cDNAs that are likely to have duplicated after the rodent-primate divergence.

Here let us explain the theory behind the numerical cutoffs of 0.6 and 1.5. The branch length between mouse and the last common ancestor (LCA) of rodents and primates is estimated to be two to three times longer than the branch length between humans and the LCA (Rat Genome Sequencing Project Consortium 2004). Here we take the larger value of 3 as the ratio and consider the situation where a gene duplicated at the same time as the rodent-primate divergence. The evolutionary distance in the neutral sites between such mouse paralogs should be $(3 + 3)/(3 + 1) = 1.5$ times that between a mouse sequence and its human ortholog. This gives one of the above cutoffs. Now, the average distance between humans and mice is estimated to be about 0.5 in

neutral sites (Mouse Genome Sequencing Consortium 2002). So the average neutral distance between the mouse paralogs considered here is expected to be 0.75, which corresponds to 0.47 differences per neutral site under the Jukes Cantor model (Jukes and Cantor 1969). The dists program of the ODEN package counts synonymous sites and synonymous differences using the method of Nei and Gojobori. This method is known to underestimate the number of synonymous sites, resulting in the overestimation of synonymous differences per synonymous site (Nei and Kumar 2000). Let us now use the transition-transversion ratio of four, an approximate average between mouse and rat orthologs. Then the correction factor is approximately 0.83. The dists program is therefore expected to give the estimation of $0.47/0.83 = 0.57$ on average for the synonymous differences per synonymous site between mouse paralogs that duplicated exactly when rodents and primates diverged. Allowing for a bit of fluctuation, we set the numerical cutoff of 0.6.

Construction and Refinement of Quartets

From each of the selected 42,546 pairs of mouse cDNAs that are likely to have diverged after the rodent-primate divergence, we tried to construct a quartet by adding the rat ortholog candidates of mouse sequences. This process also filters out mouse cDNA pairs that appear to have duplicated after the mouse-rat speciation.

When constructing quartets, we imposed the following conditions:(1) each member of the mouse pair has at least one rat ortholog candidate and (2) the rat ortholog candidates of the mouse sequences are transcribed from the genes distinct from each other. By applying these two criteria, we succeeded in constructing quartets for 6,568 mouse pairs. However, some of these 6,568 quartets contain mouse cDNA pairs that are transcribed from the same gene pair (and similar situations can also occur for rat gene pairs). We thus chose 3,657 "representative quartets," each of whose mouse cDNA pair has the highest alignment score among the alternative splicing variants of a gene pair.

We constructed multiple alignments of cDNAs for those 3,657 quartets in the manner described below. Then we inferred phylogenetic relationships among the sequences in each quartet by using the neighbor-joining method with a distance matrix estimated by Kimura's two-parameter model. Finally, we selected 2,641 quartets, each of which contains two orthologous mouse-rat pairs and whose phylogenetic trees indicate that duplication events occurred before the mouse-rat divergence. The preceding screening already narrowed down the duplication events to after the primate-rodent divergence. These 2,641 quartets are the basis for the later analyses.

Construction of Multiple Alignments of cDNA Sequences

Here we explain how we constructed multiple alignments of cDNA sequences used in this study. Given a set of cDNA sequences, we began by constructing a multiple alignment of their peptide counterparts via the protein mode of ClustalW (Thompson, Higgins, and Gibson 1994) version 1.83 with the default parameters. Then we constructed its corresponding codon alignment guided by the correspondence between cDNAs and peptides fetched from the EnsEMBL database. The process up to here is a common practice for constructing a multiple alignment of codon sequences, which can be seen in papers on codon sequence analyses. However, we encountered quite a few cases in which the alignment constructed this way looks erroneous due to frameshifts in short intervals or misplacements of long insertions/deletions. Such alignment errors might be fatal when we try to detect gene conversion because it might cause numerous false positives. We therefore devised a method to remove potential alignment errors by comparing three multiple alignments constructed from the same sequence set but under different parameter settings. The two alignments to be compared to the above codon alignment were produced by applying the nucleotide mode of ClustalW to this codon alignment under two parameter sets: (dnamatrix, gapopen, gapext) = (ClustalW, 4, 0) and (ClustalW, 12, 0). Finally, we masked regions in the codon alignment that showed discrepancies with at least one of the two nucleotide alignments constructed as above, thus maintaining only regions that are robust under moderate parameter changes. When actually conducting analyses, we discarded gene sets that had lost a large fraction of alignments (30% in this study) through this masking process.

Statistical Tests to Detect Gene Conversion

Here let us briefly explain our strategy for detecting gene conversion. We are considering the situation where a gene duplication event precedes the speciation of mouse and rat, producing a quartet of genes: mouse1, rat1, mouse2, and rat2, where mouseα and ratα are orthologous ($\alpha = 1$ or 2), while genes 1 and 2 are paralogous (fig. 1). If base substitutions dominate the sequence evolution, the major informative sites should be of "type 1," clustering orthologous sequences (fig. 1*A*). If gene conversion occurs, however, we expect a significantly large number of a second type of informative sites, the "type 2" sites, which cluster paralogous sequences in the same species (fig. 1*B*). It should be noted that simple parallel substitutions can also generate type 2 sites besides a third type of informative sites, the "type 3" sites (fig. 1*C*). Therefore, in order for the observed type 2 sites to indicate gene conversion, they should be significantly more abundant than expected by parallel substitutions alone. Besides, gene conversion will be further supported if the type 2 sites are segregated from other types of informative sites along the multiple alignment of a quartet.

In order to examine whether these two conditions are satisfied or not, we conducted the following four statistical tests: (1) a test for the count of type 2 informative sites (the IScomp test), which turned out to be similar in spirit to but technically different from the codouble method of Balding, Nicholas, and Hunt (1992); (2) a test for the size of the longest run of type 2 sites (the T2run test), which is mathematically similar to Stephens' (1985) test; (3) a test for the number of consecutive runs of informative sites of the same types (the SameTrun test), which is Takahata's (1993) two-sample runs test applied to a slightly different situation; and (4) a test for the regional variation in sequence

similarities (the CSrun test), exploiting the GENECONV software (Sawyer 1989, http://www.math.wustl.edu/~sawyer). Then, with the aid of computer simulations, we integrated the results of those four tests in order to achieve high sensitivity for gene conversion detection while keeping low false-positive rate. In order to let the readers have a feeling on our detection method, we have prepared Supplementary Figure S1*A–C* (Supplementary Material online). The figure exemplifies the results of our statistical tests conducted on three quartets, one "extremely positive," one "moderately positive," and one "gray." Details of the statistical tests, computer simulations, and integration procedures are described in Supplementary File (Supplementary Material online). There the readers will also find tests of our method's performance and comparisons of the performance between our method and GENECONV.

Inference of Mouse cDNA Pairs That Have Undergone Gene Conversion

Because the type 2 sites do not tell which of the mouse pair and rat pair was affected by gene conversion, we used GENECONV (in the NUCL mode) and synonymous substitutions in order to infer the pairs struck by gene conversion. We examined each quartet showing signs of gene conversion, and we judged that gene conversion hits the mouse pair if either of the following conditions was satisfied: (1) GENECONV detected a tract of $P < 0.05$ in the mouse pair or (2) the best putative gene conversion tract in the mouse pair had a synonymous distance significantly smaller than that between orthologs. We used the binomial test and the threshold of 0.1. The resulting positive mouse cDNA pairs were used for the correlation analyses.

Correlations of Gene Conversion Susceptibilities with Properties of Gene Pairs

From the EnsEMBL FTP site (ftp.ensembl.org/pub), we fetched mapping of exons onto chromosomes as well as information on exon components, gene origins, and translation start and end points of transcripts. Integration of these pieces of information yielded the mapping of peptide-coding regions onto the chromosomes. Using this mapping, we first classified gene pairs according to the linkages, namely, whether the two genes reside on the same chromosome or not. Linked gene pairs were further classified according to their relative transcriptional orientations into three categories: "head-to-tail" (5′-3′ 5′-3′), "head-to-head" (3′-5′ 5′-3′), and "tail-to-tail" (5′-3′ 3′-5′). We also classified the quartets according to physical distances. Here the physical distance of a pair was defined as the number of base pairs between the coding regions of the cDNA sequences mapped on a chromosome. When examining the correlations of gene conversion prevalence with relative orientations and with physical distances, we used only those quartets whose mouse and rat pairs have the same relative orientation because such quartets must have rarely, if any, drastically changed their orientations or physical distances after the mouse-rat speciation. Because we collected quartets in a mouse-centered manner, we only used the distance between mouse sequences in each quartet. Then we sorted the quartets in ascending order of distance and distributed them into four bins containing almost identical numbers of quartets and counted "positive" mouse cDNA pairs in each of the four bins. Then we examined whether there are correlations between these classifications and the proportion of positive mouse pairs.

In similar manners, we examined correlations of the incidence of gene conversion with other properties of genes, such as the peptide length, exon count, synonymous differences per synonymous site, degree of bootstrap support, and size of the gene family the mouse pair belongs to. Here we defined a "family" as a set of genes that are inferred to have diverged from each other after the rodent-primate divergence. We then categorized the gene pairs by the size of the "subfamily," which is defined as a cluster of family members contiguous to each other along the chromosome.

We also examined the dependence on the functional categories. We performed InterProScan (Zdobnov and Apweiller 2001; http://www.ebi.ac.uk/interpro) version 4.0 on mouse peptides whose cDNA counterparts belong to the 2,641 quartets. We then extracted domain candidates whose exact ID numbers are shared by both sequences in the mouse pair. And we kept only those domains each of whose regions in the two mouse sequences overlap each other in the alignment. We counted the "positive," "gray," and "negative" quartets in each of the categories belonging to the InterPro domain or family, as well as in each of the gene ontology terms (The Gene Ontology Consortium 2000; http://geneontology.org) whenever defined. Both upper tailed and lower tailed *P* values are calculated by using Fisher's exact test.

In Silico Verifications of the Robustness of Our Results

Because some of the methods we employed in this study are not yet widely accepted, some anxieties might remain regarding whether the results obtained here are biologically relevant or there are some artifacts that grossly deform the picture. We therefore conducted three independent additional analyses using (1) the Homologous Vertebrate Genes (HOVERGEN) database (Duret, Mouchiroud, and Gouy 1994), (2) a set of "syntenic" ortholog candidates, and (3) the simpler, "standard," alignment construction.

(1) Although well motivated and planned, our series of screening process to sample quartets is a bit complicated and different from standard methods of homologous sequence collection. A concern may therefore arise that some artifacts might have infiltrated the process. Thus, we reanalyzed the incidence of gene conversion in a data set extracted from the HOVERGEN database (Duret, Mouchiroud, and Gouy 1994; http://pbil.univ-lyon1.fr/databases/hovergen.html) release 47. The database contained 49,206 mouse sequences (including redundancy) distributed in 11,776 families. We first examined the phylogenetic trees and retrieved 1,901 subfamilies of mouse sequences that were inferred to have been formed after the mammalian radiation. Because the current version of HOVERGEN contains redundant sequences (see the above Web site), we removed the redundancy by extracting only those mouse entries that have links to EnsEMBL. This left us with 128

subfamilies, containing 436 nonredundant EnsEMBL mouse gene pairs. A total of 159 pairs successfully formed quartets. Finally, we chose 134 quartets whose phylogenetic trees suggested that the duplication events took place before the mouse-rat speciation. Then we applied our gene conversion search to these quartets.

(2) Although our method to retrieve rat ortholog candidates are close to standard, sequence similarity alone might not be enough to infer the orthology when we handle multiple-gene families. We therefore prepared a more stringent set of ortholog candidates by imposing "conserved synteny" as an additional condition. Our operational definition of conserved synteny for a pair of mouse and rat ortholog candidates is the following: (1) the mouse gene has at least a gene that lies within 1 Mb of it and has different ortholog candidates from those of the gene in question; (2) when taking the closest of such genes on each side, its ortholog candidate is also within 1 Mb of the ortholog candidate of the gene in question; (3) they have a conserved gene order relative to the transcriptional direction of the gene in question; and (4) when the gene have such closest genes on both sides, both must satisfy the conditions (2) and (3). We finally selected such syntenic ortholog candidates for 12,341 mouse genes. Then we constructed 676 refined quartets, each consisting of two syntenic mouse-rat ortholog candidates that are inferred to have diverged between the rodent-primate divergence and the mouse-rat speciation. Using these 676 refined quartets, we reanalyzed the incidence of gene conversion, as well as correlations between the proportion of positive gene pairs and various properties of genes. We put the results of statistical analyses on this refined set of quartets into Supplementary Tables S1 through S9, S12, and S13 (Supplementary Material online).

(3) Although our alignment-masking process was originally devised to reduce false-positive rate, it would be desirable to check whether the masking process really works well. We therefore repeated our analysis using simpler, standard, codon alignments, which were produced by just replacing amino acids in the protein alignments with their codon counterparts. We constructed such alignments for 3,657 quartets that we had already prepared, and 2,582 of them satisfied the condition on the inferred tree topology. Then we compared the result of gene conversion detection on these standard alignments with that obtained by using our masked alignments.

We also confirmed the credibility of our gene conversion detection methods by conducting performance tests on data sets generated by computer simulations. The Supplementary File (Supplementary Material online) describes these analyses, as well as the comparison of our method with GENECONV.

Results

Number of Quartets Affected by Gene Conversion

Based on the result of our statistical test, we classified our 2,641 quartets into four categories: extremely positive, moderately positive, gray, and negative (see Supplementary File for details, Supplementary Material online). Roughly speaking, an extremely positive quartet is defined as a quartet with the simulated P value under 0.0001, and a moderately positive quartet has the P value under 0.005. The results of computer simulations suggested that the former and the latter sets should contain less than two and less than 19 false positives out of the 2,641 quartets, respectively, in terms of 95% one-tailed confidence interval (Supplementary File, Supplementary Material online). We also divided nonpositive categories into gray and negative. A gray status was assigned to quartets that are difficult to judge whether gene conversion has occurred or not. The simulated P value of 0.125 was chosen as the boundary between gray and negative in this study.

Table 1
Summary of Statistical Tests Conducted on the Mouse-Rat Quartets as well as on the Mouse and Rat cDNA Pairs

Test Status	Quartets	Mouse_Pairs[a]	Rat_Pairs[b]
Positive1[c]	244 (9.2%)[d]	151 (5.7%)	153 (5.8%)
Positive2[e]	244 (9.2%)	189 (7.2%)	125 (4.7%)
Gray[f]	773 (29.3%)	770 (29.2%)	810 (33.4%)
Negative	1,380 (52.3%)	1,531 (58.0%)	1,546 (48.2%)
Total	2,641 (100%)	2,641 (100%)	2,641 (100%)

[a] Results of tests on mouse cDNA pairs contained in quartets.

[b] Results of tests on rat cDNA pairs contained in quartets. Because of redundancy, the figures in this column do not necessarily reflect the real situation. In the nonredundant set of rat pairs, the proportions of Positive1 and Positive2 are 10.1% (102/1,009) and 8.3% (84/1,009), respectively.

[c] The extremely positive set consisting of quartets that showed almost unequivocal signs of gene conversion (approximately equivalent to the condition: $P < 0.0001$).

[d] The number on the left of parentheses and that in parentheses are the count and the proportion, respectively, of quartets or mouse pairs with particular test status.

[e] The moderately positive set consisting of quartets whose test results indicated the occurrences of gene conversion (approximately equivalent to the condition: $P < 0.005$).

[f] Sets of quartets that are difficult to judge whether gene conversion affected them or not (approximately equivalent to the condition: $P < 0.125$).

Out of the 2,641 quartets we examined, 244 (ca. 9%) were classified as extremely positive and 244 more (ca. 9%) were classified as moderately positive for gene conversion (table 1). In total, we detected significant signs of gene conversion (at the simulated false-positive rate 0.5%) in 488 quartets, which is about 18% of the sample size. We have to note that the gray and negative categories could also contain quite a few quartets affected by gene conversion. But the exact number of such quartets can vary enormously depending on the actual modes of their evolution via substitutions. However, if we naively believe our simulation results that approximately 20% of true-positive signs escaped our detection method (Supplementary File, Supplementary Material online), about 120 more quartets that experienced gene conversion should be hidden in the gray and negative categories, giving the estimate that about 610 (ca. 23%) of the 2,641 quartets underwent gene conversion.

We also classified mouse cDNA pairs in a similar manner and found 151 pairs (ca. 6%) as extremely positive and 189 (ca. 7%) as moderately positive (table 1). The total number of positive mouse pairs are therefore 340 (ca. 13%). Combining these results with the result on rat, 130 quartets showed highly significant signs of gene conversion in both mouse and rat gene pairs.

Table 2
Counts and Proportions of Positive, Gray, and Negative Mouse Pairs Classified by the Linkage

Linkage[a]	Both	Mouse Only	Rat Only	Neither	Unknown	Total
Positive	271 (15.1%)[b]	8 (17.8%)	0 (0%)	2 (1.1%)	59 (11.0%)	340 (12.9%)
Gray[c]	551 (30.6%)	15 (33.3%)	14 (17.3%)	31 (17.0%)	159 (29.8%)	770 (29.2%)
Negative	977 (54.3%)	22 (48.9%)	67 (82.7%)	149 (81.9%)	316 (59.2%)	1,531 (58.0%)
Total	1,799	45	81	182	534	2,641

[a] The key for linkage categories—both: both mouse and rat pairs are linked, neither: neither mouse nor rat pair is linked, unknown: the linkage of either mouse or rat pair is unknown.

[b] The number in parentheses is the proportion of mouse cDNA pairs showing a particular test status in the set of mouse pairs with a particular linkage category.

[c] Mouse pairs that are difficult to judge whether gene conversion affected them or not.

Correlations with Linkage, Physical Distance, and Relative Orientation

We examined how the proportion of positive mouse cDNA pairs varies depending on their relative positional properties.

Most of the collected mouse gene pairs consisted of genes that were linked (residing on the same chromosome): linked pairs accounted for 87.5% (=1,844/2,107) of the pairs with known linkage status (table 2). The table also indicates that linked gene pairs (279/1,844 = 15%) are significantly more prone to gene conversion than unlinked ones (2/263 = 0.8%, $P = 3.0 \times 10^{-8}$ by Fisher's exact test).

To see the dependence on physical distance, we first sorted the quartets in ascending order of the distance between mouse sequences, and then we distributed the quartets into four bins of almost equal sizes (table 3). The table shows a clear negative correlation between the proportion of positive pairs and the physical distance ($P = 1.2 \times 10^{-9}$ by the chi-square test of df = 3). While as much as 29% of mouse pairs within 55 kb were positive for gene conversion, only 10% of those over 371 kb showed signs of gene conversion (table 3). To see the distance dependence in more detail, we subdivided the closest and remotest categories into smaller bins. Mouse pairs within 27 kb were even more prone to gene conversion, with the frequency of 36% significantly higher than that of 23% for pairs between 27 and 55 kb (table 3, $P = 0.011$ in Fisher's exact test). For the remote gene pairs, however, the frequency hovered between 7% and 20% and did not fall as the distance increased (table 3).

Because the sample sizes are small, we do not know whether the frequency of 20% represents the proper biology for these remote genes or not, but at least some gene pairs with their distances over 10 Mb were positive for gene conversion. This might be at least partially attributable to recent chromosomal remodeling such as translocations and inversions. To support this idea, when we conducted the same analysis on the refined set of quartets consisting only of orthologous pairs of conserved synteny, we did not find signs of gene conversion among the 34 pairs over 811 kb (Supplementary Table S3, Supplementary Material online).

Next we classified the pairs into three categories of relative transcriptional orientation: head-to-tail (5′-3′ 5′-3′), head-to-head (3′-5′ 5′-3′), and tail-to-tail (5′-3′ 3′-5′). The proportion of positive mouse pairs does not depend so much on the relative orientations (table 4). Although the head-to-head set may appear more abundant in positive pairs, the bias is not significant ($P > 0.2$ by Fisher's exact test). We also examined dependence on the relative orientation for each of the four categories of the physical distance (Supplementary table S4*B*,*C*, Supplementary Material online). In most distance categories, gene conversion frequency appeared to vary among relative orientations only within reasonable sampling fluctuations. Only the third

Table 3
Counts of Positive, Gray, and Negative Mouse Pairs Classified by Their Physical Distances

Distance	Positive	Gray[a]	Negative	Total
0–55 kb	89 (29.3%)[b]	99 (32.6%)	116 (38.2%)	304
55–167 kb	48 (15.8%)	97 (32.0%)	158 (52.1%)	303
167–371 kb	46 (15.1%)	94 (30.9%)	164 (53.9%)	304
371–90 Mb	30 (9.9%)	87 (28.6%)	187 (61.5%)	304
Total	213 (17.5%)	377 (31.0%)	625 (51.4%)	1,215
Subdivision of the closest category				
0–27 kb	54 (35.5%)	48 (31.6%)	50 (32.9%)	152
27–55 kb	35 (23.0%)	51 (33.6%)	66 (43.4%)	152
Subdivision of the remotest category				
371–811 kb	11 (7.2%)	46 (30.3%)	95 (62.5%)	152
811–1.6 Mb	5 (6.6%)	26 (34.2%)	45 (59.2%)	76
1.6–9.5 Mb	7 (18.4%)	6 (15.8%)	25 (65.8%)	38
9.5–25 Mb	2 (10.5%)	4 (21.1%)	13 (68.4%)	19
25–90 Mb	5 (26.3%)	5 (26.3%)	9 (47.4%)	19

NOTE.—We used only quartets whose mouse and rat pairs have the same relative orientation.

[a] Mouse pairs that are difficult to judge whether gene conversion affected them or not.

[b] The figure in parentheses is the proportion of mouse pairs showing a particular test status in a particular distance class.

Table 4
Counts and Proportions of Positive, Gray, and Negative Mouse Pairs Classified by Their Relative Orientations

Relative Orientation[a]	Head-to-Tail	Head-to-Head	Tail-to-Tail	Total
Positive	150 (17.4%)[b]	31 (19.6%)	32 (16.6%)	213 (17.5%)
Gray[c]	284 (32.9%)	43 (27.2%)	50 (25.9%)	377 (31.0%)
Negative	430 (49.8%)	84 (53.2%)	111 (57.5%)	625 (51.4%)
Total	864	158	193	1,215

NOTE.—We used only quartets in which mouse and rat pairs have the same relative orientation.

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

[b] The figure in parentheses is the proportion of mouse pairs showing a particular test status in the set of mouse pairs with a particular relative orientation.

[c] Mouse pairs that are difficult to judge whether gene conversion affected them or not.

category of distance between 167 and 371 kb showed an interesting behavior, where gene conversion favored pairs of opposite orientations, with the frequency of 21% (=22/105) for opposite orientations and 12% (=24/199) for the same orientation ($P = 0.031$ in Fisher's exact test). But this bias lost the statistical significance in the refined set of syntenic quartets ($P = 0.27$), probably due to the small sample size. It remains to be seen whether this bias is biologically, or evolutionarily, significant or not.

Correlations with Family Sizes, Synonymous Differences, and Other Physical and Evolutionary Parameters

We analyzed how the susceptibility to gene conversion depends on the sizes of the family and subfamily the gene pair belongs to. Here we define a family as a set of genes that have been generated by gene duplication events after the rodent-primate divergence. A subfamily is defined as a cluster of physically adjacent genes in a family. Both the dependences showed very similar behaviors to each other. Because we have seen a clear negative correlation between the gene conversion susceptibility and the physical distance in the last subsection, a subfamily seems to be the more relevant unit than a family when discussing gene conversion. We therefore focused on subfamilies.

Table 5
Counts of Positive, Gray, and Negative Mouse Pairs Classified by the Sizes of Subfamilies They Belong to

Subfamily Size (n)	Positive	Gray[a]	Negative	Total
Different[b]	192 (10.4%)[c]	520 (28.1%)	1,136 (61.5%)	1,848
$n = 2$	19 (27.5%)	27 (39.1%)	23 (33.3%)	69
$n = 3$	25 (38.5%)	20 (30.8%)	20 (30.8%)	65
$n = 4$	12 (32.4%)	15 (40.5%)	10 (27.0%)	37
$n = 5$	25 (35.7%)	18 (25.7%)	27 (38.6%)	70
$n = 6$	7 (21.9%)	12 (37.5%)	13 (40.6%)	32
$n = 7$	8 (16.3%)	24 (49.0%)	17 (34.7%)	49
$n = 8, 9$	20 (13.0%)	63 (40.9%)	71 (46.1%)	154
$10 <= n < 15$	31 (13.2%)	57 (24.4%)	146 (62.4%)	234
$15 <= n < 20$	1 (1.2%)	14 (16.9%)	68 (81.9%)	83
Same[d]	148 (18.7%)	250 (31.5%)	395 (49.8%)	793
Total	340 (12.9%)	770 (29.2%)	1,531 (58.0%)	2,641

NOTE.—Here we define a subfamily as a cluster of contiguously located mouse genes that are inferred to have diverged from one another after the rodent-primate divergence.

[a] Mouse pairs that are difficult to judge whether gene conversion affected them or not.

[b] The numbers of mouse pairs whose member genes belong to different subfamilies.

[c] The figure in parentheses is the proportion of mouse pairs showing a particular test status in a particular class of family sizes.

[d] The subtotal numbers of mouse pairs whose member genes belong to the same subfamily.

Table 6
Counts of Positive, Gray, and Negative Mouse Pairs Classified by the Proportional Synonymous Differences Between the Member Genes

Prop. sdiff[a] (Ps)	Positive	Gray[b]	Negative	Total
$0.1 \leq Ps < 0.2$	56 (32.9%)[c]	57 (33.5%)	57 (33.5%)	170
$0.2 \leq Ps < 0.3$	175 (21.9%)	292 (36.5%)	332 (41.6%)	799
$0.3 \leq Ps < 0.4$	77 (14.2%)	174 (32.1%)	291 (53.7%)	542
$0.4 \leq Ps < 0.5$	29 (4.6%)	143 (22.7%)	457 (72.7%)	629
$0.5 \leq Ps < 0.6$	3 (0.6%)	104 (20.8%)	394 (78.6%)	501
Total	340 (12.9%)	770 (29.2%)	1,531 (58.0%)	2,641

[a] The proportion of synonymous differences (Prop. sdiff), which means synonymous differences per synonymous site of the entire alignment between the two mouse sequences in a pair.

[b] Mouse pairs that are difficult to judge whether gene conversion affected them or not.

[c] The figure in parentheses is the proportion of mouse pairs showing a particular test status in a particular class of synonymous differences.

Gene pairs belonging to smaller subfamilies are more prone to gene conversion (table 5). For example, pairs belonging to subfamilies of size less than seven show the gene conversion prevalence of 32% (=88/273), while the prevalence for subfamilies of size seven or more is 12% (=60/520), showing an enormous statistical significance ($P = 3.6 \times 10^{-12}$ in Fisher's exact test). Another interesting point is that "isolated" gene pairs, which are equivalent to subfamilies of size two, look less prone to gene conversion than subfamilies of size between three and five, although without statistical significance ($P = 0.12$ between sizes two and three). These observations should help model the evolution of gene families under the influence of gene conversion.

Table 6 shows the dependence of gene conversion frequency on the number of synonymous differences per synonymous site among mouse paralogous sequences in an entire quartet alignment. It is obvious from the table that gene conversion is more prevalent among pairs with higher sequence similarities. It is unclear, however, whether the high similarity accelerated gene conversion or frequent gene conversion maintained the high similarity. Probably, both mechanisms worked together to form a kind of positive feedback loop. To elucidate this relationship, however, a more thorough data analysis would be necessary, maybe with the aid of some models.

We also examined the dependence of gene conversion incidence on peptide lengths, exon numbers, and bootstrap support for the clustering of mouse-rat orthologous pairs. However, we did not find any trends that can be simply interpreted. So we just presented the results in Supplementary Tables S7–S9 (Supplementary Material online). It remains to be seen whether some trend will be revealed or not when we incorporate correlations among these parameters as well as with others.

Table 7
Counts of Total and Positive Mouse Pairs in Functional Categories (excerpt)

Interpro ID (name or description)	Total	Positive	Pv_Lower[a]	Pv_Upper[b]
Three major families				
IPR000276[c] (rhodopsin-like GPCR receptor)	1,465	200	0.838*	0.193
IPR000725 (olfactory receptor)	666	120	1.00	1.52×10^{-5}
IPR004072 (vomeronasal receptor, type 1)	402	23	1.72×10^{-7}*	1.00
Rich in gene conversion				
IPR002126 (cadherin)	35	25	1.00	2.08×10^{-15}*
IPR009072 (histone-fold)	21	11	1.00	1.75×10^{-5}
IPR001664 and IPR011000[d] (intermediate filament protein and apolipophorin III like)	6	5	1.00	2.02×10^{-4}*
IPR001254 (peptidase S1, chymotrypsin)	21	9	1.00	7.12×10^{-4}
IPR002957 (keratin, type I)	5	4	1.00	1.30×10^{-3}*
IPR000379 (esterase/lipase/thioesterase)	30	10	9.99×10^{-1}	3.48×10^{-3}
IPR003597 and IPR001039[d] (immunoglobulin C1 type and major histocompatibility complex protein, class I)	17	7	9.99×10^{-1}	3.75×10^{-3}
IPR002018[e] (carboxylesterase, type B)	28	9	1.00	7.22×10^{-3}
IPR006862 (Acyl-CoA thioester hydrolase/bile acid-CoA amino acid *N*-acetyltransferase)	2	2	1.00	1.72×10^{-2}*
IPR003439 and IPR003593[d] (ABC transporter related and AAA ATPase)	2	2	1.00	1.72×10^{-2}*
IPR002213 (UDP-glucuronosyl/ UDP-glucosyltransferase)	13	3	0.921	0.238*
Poor in gene conversion				
IPR004073[f] (vomeronasal receptor, type 2)	279	0	8.06×10^{-19}	1.00
IPR000337 (GPCR family 3, metabotropic glutamate receptor like)	282	1	2.47×10^{-17}	1.00
IPR001500 (nine cysteins of GPCR)	203	1	3.97×10^{-12}	1.00
IPR001828 (extracellular ligand binding receptor)	160	1	2.13×10^{-9}	1.00
IPR003596 (immunoglobulin V type)	252	10	3.06×10^{-7}	1.00
IPR007110 (immunoglobulin like)	246	17	8.90E-04	1.00

NOTE.—ABC, ATP-binding cassette; ATPase, adenosine triphosphatase; GPCR, G-protein-coupled receptor; UDP, uridine diphosphate. The numbers and *P* values in this table are estimated using the whole set of quartets, an asterisk indicates the significance when estimated with the refined set of syntenic quartets. You can find complete tables as Supplementary Tables S10*A*,*B* and S11*A*-*C* for the analysis using the set of all quartets and S12*A*,*B* and S13*A*-*C* for that with the set of syntenic quartets (Supplementary Material online).

[a] The lower tailed *P* value, which is the probability that there are positive pairs less than or equal to those observed under the null hypothesis of the even distribution.

[b] The upper tailed *P* value, which is the probability that there are positive pairs more than or equal to those observed under the null hypothesis of the even distribution.

[c] This family is a composite superfamily consisting of the other two major families, that is, olfactory receptors and vomeronasal receptors, as well as some others.

[d] These two domains always occurred in pair.

[e] This domain was always accompanied by IPR000379 (esterase/lipase/thioesterase), which is two entries above.

[f] This family almost completely overlapped IRP000337 (GPCR family 3), which is just below.

Correlation with Functional Categories

In order to see the differences in susceptibilities among different functional categories, we classified mouse gene pairs according to the functional domains shared by the member sequences (table 7; Supplementary Tables S10*A*,*B* and S11*A*–*C*, Supplementary Material online). The prevalence of gene conversion did vary across functional categories. In terms of the total number of gene pairs, the commonest category is the rhodopsin-like G-protein-coupled receptor (GPCR) superfamily (1,465 pairs). It consists of many gene families including the second and third major ones, namely, the olfactory receptors (666 pairs) and the vomeronasal receptors type 1 (402 pairs). These two families substantially varied in their susceptibility to gene conversion: the olfactory receptors were far more susceptible than normal (120 positives, $P = 1.5 \times 10^{-5}$) and the vomeronasal receptors were extremely immune to gene conversion (23 positives, $P = 1.7 \times 10^{-7}$). Many families and domains were significantly rich in gene conversion, and some were significantly poor in gene conversion (table 7). Some of the categories lost the statistical significance when we reanalyzed the functional dependence using the refined set of syntenic quartets (table 7; Supplementary Tables S12*A*,*B* and S13*A*–*C*, Supplementary Material online). This is partly because the refined set is almost free from overrepresentation due to recent duplication events.

We further delved into the categories whose statistical significance remained after using the refined set. We reexamined, for each functional categories, correlations of the frequency of gene conversion with the physical distance, subfamily size, and synonymous differences per synonymous site. The previous subsections have already shown that these three quantities have remarkable correlations with the average prevalence of gene conversion. First, we compared the two major families the olfactory receptors and the vomeronasal receptors to see what makes these two

functionally similar families so differently susceptible to gene conversion (Supplementary Table S14*A*,*B*, Supplementary Material online). The olfactory receptor family shows almost average behavior (Supplementary Table S14*A*, Supplementary Material online). Its enhanced incidence of gene conversion seems to be attributable to its properties, such as no unlinked pairs, and a slightly high proportion of pairs each embedded in a subfamily. On the other hand, the low gene conversion incidence of the vomeronasal receptor family is hard to explain with the physical properties alone. It abounds with linked pairs of the remotest category, and it is poor in pairs belonging to the subfamilies of size five or less. However, it showed low frequency even in categories that are normally prone to gene conversion, such as that of physical distance within 55 kb. Although the final conclusion would need the correlation analysis of the three quantities, some special biological mechanism might have hampered gene conversion on the vomeronasal receptor family.

We next examined the families of cadherins, intermediate filament proteins, and keratins (Supplementary Table S14*C*–*E*, Supplementary Material online), which are significantly prone to gene conversion (table 7). Intermediate filament proteins and keratins showed ideal combinations of physical properties (Supplementary Table S14*D*,*E*, Supplementary Material online): small distances within pairs, contiguous cluster sizes between three and five, and small synonymous differences per synonymous site. Because they contain only a small numbers of pairs, their high gene conversion incidence are well accounted for by this "lucky combination" of favorable conditions. The cadherin family, however, does not look perfect (Supplementary Tables S14*C*, Supplementary Material online). The pair distances fall within 167 kb, and the pairs belong to contiguous clusters that are either small, with sizes between three and five, or large, with sizes between 10 and 14. And their synonymous differences are not necessarily small enough. These observations indicate that cadherins might have been under some selective pressure that favors the occurrence of gene conversion. We also looked into the functional categories of "nine cysteins of GPCR" and "immunoglobulin V type" (Supplementary Table S14*F*,*G*, Supplementary Material online), which are significantly immune to gene conversion (when using the whole set of quartets, table 7). The former category turned out to be endowed with the "worst combination" of conditions that disfavor gene conversion: most of the gene pairs are unlinked or in the remotest category, have two members belonging to different contiguous clusters, and have large synonymous differences per synonymous site. The category of immunoglobulin V type is harder to interpret. Although it abounds with physically remotest pairs (and those with "unknown" linkage), it also contains quite a few pairs in physically closer categories. And it is rich in pairs with small synonymous differences. It seems difficult for the physical properties alone to explain this paucity of gene conversion in this family, which is suggestive of some biology underneath the evolution of this family.

In Silico Verifications of the Robustness of Our Results

In order to alleviate the concern that our data preparation might entail some artifacts, we checked the robustness of our main results by comparing them with those of independent analyses conducted with three different sets of input data. The three input data sets were prepared by using (1) the HOVERGEN database (Duret, Mouchiroud, and Gouy 1994), (2) a set of syntenic ortholog candidates, and (3) the simpler, standard, alignment construction.

Supplementary Table S15*A*–*C* (Supplementary Material online) summarizes the results of these comparisons. As we can see, (1) the quartets constructed from HOVERGEN entries give positive quartets whose proportion of 22% (=29/134) is similar to that obtained from our set of quartets (Supplementary Table S15*A*, Supplementary Material online; table 1); (2) the refined set of 676 syntenic quartets contains a larger proportion of positive quartets (175/676 = 26%) than our whole set does (488/2,641 = 18%, Supplementary Table S15*B*, Supplementary Material online), which should be attributable to the enrichment of quartets with conserved orientation (Supplementary Table S2, Supplementary Material online; tables 2 and 4); and (3) our masked alignments yield more stringent results of statistical tests than the standard alignments do (Supplementary Table S15*C*, Supplementary Material online), suggesting that our masking method did reduce false positives.

Comparison (3) revealed 19 quartets that were positive when using masked alignments but gray when using standard alignments. Inspection of the alignments showed that each putative gene conversion tract straddled a masked region that contains type 1 sites. Because it was difficult to judge whether these positive calls were false or authentic, we reassigned the gray status to these 19 quartets. This reduced the number of positive quartets from 507 to 488. This reassignment of the status has already been, and will always be, reflected in *Results* and *Discussion* except this subsection.

We also repeated our analyses using the refined set of syntenic quartets. We did not find remarkable differences from the result using the whole set. Because some people may think it better to use this set of syntenic quartets, we presented the results with this set in Supplementary Tables S1 through S9, S12, and S13. Supplementary Tables S1 through S6 correspond to tables 1 through 6, respectively. And Supplementary Tables S12 and S13 correspond to Supplementary Tables S10 and S11, respectively (Supplementary Material online).

Discussion

Genomic Prevalence of Gene Conversion?

Out of the 2,641 mouse cDNA pairs collected, 340 pairs (ca. 13%) showed significantly positive signs of gene conversion, with at most 20 false positives expected. This means that gene conversion is far from a rarity but rather is ubiquitous across a mammalian genome. So there is good reason that we should be cautious when inferring the duplication date and the phylogenetic relationships of duplicated genes.

We note that we examined only those mouse pairs each of which is included in a quartet with the inferred phylogenetic relationship of ((mouse1, rat1), (mouse2, rat2)). Thus, our subject mouse pairs are expected to have resulted from duplication events postdating the divergence of rodents and primates and predating the speciation of mouse and rat. In

this study, in order to reduce the "contamination" by false positives, we deliberately dismissed about 40,000 mouse cDNA pairs. These dismissed pairs are divided into two broad categories: (1) mouse gene pairs that appear to have duplicated after the mouse-rat speciation and (2) those that lack rat ortholog candidates. Many studies report the positive correlation between gene conversion frequency and the sequence similarity (Liskay, Letsou, and Stachelek 1987; Elliott et al. 1998; Lukacsovich and Waldman 1999; Semple and Wolfe 1999), which is consistent with our result (table 6). It would be therefore natural to conjecture that an enormous number of gene conversion events should be lurking in those unexamined pairs, especially in category (1). We expect that gene conversion should be more common among those gene pairs, pushing the overall prevalence of gene conversion much higher than the current estimate of 15%.

In order to see whether this is indeed the case or not, we have to establish a detection system that suppresses false positives to a low enough level and yet can efficiently detect multiple gene conversion tracts. Such multiple tracts tend to foil any existing detection methods and are therefore almost indistinguishable from a recent duplication event. Probably, inclusion of introns and flanking regions should be conducive to the effective detection of multiple gene conversion events. This is because introns and flanking regions tend to evolve faster than coding regions, giving enough background information, such as the type 1 sites, that highlights signs of gene conversion even when coding regions are very similar to each other. This study, on the other hand, focused on the mouse gene pairs whose coding regions have diverged sufficiently far from each other. So they should give enough background information to allow highly sensitive detection of gene conversion. Incorporating intron sequences into this study would rather have deteriorated the quality of the analyses because of the difficulty in aligning introns, which may have experienced frequent genomic remodeling.

Comparisons with the Previous Genome-Wide Searches for Gene Conversion

In the past, genome-wide searches for gene conversion were performed on nematode worm *Caenorhabditis elegans* (Semple and Wolfe 1999) and yeast *Saccharomyces cerevisiae* (Drouin 2002). We compared our results on the mouse genome with those previous genome-wide studies.

In the present study, we detected putative gene conversions in about 13% (340/2,641) of mouse pairs examined. This figure is about two times larger than in yeast (7.8% = 69/879) and about seven times larger than in worm (2% = 143/7,829).

Some of these differences are attributable to the different nature of the collections of gene pairs. The set of mouse gene pairs we examined abounds with linked pairs (1,844 = 88% out of the 2,107 pairs with known linkage states), which are a minority in both sets of yeast and worm gene pairs (41/879 = 4.7% and 3,347/7,829 = 43%, respectively). We therefore estimated the proportions of positive pairs separately for linked and unlinked categories. For linked pairs, the figures are 15% (279/1,844) for mouse, 39% (16/41) for yeast, and 3% (104/3,347) for worm. For unlinked pairs, they became 0.8% (2/263) for mouse, 6.3% (53/838) for yeast, and 0.9% (39/4,482) for worm. In both categories, yeast stands out in the proportions of positive pairs. One conspicuous feature is that the linked pairs in worm appear poor in gene conversion. Although differences in the sequence similarity spectrums may explain these observations at least partially, we could not find enough data on yeast and worm to correct for the effect of the sequence similarity. So it remains obscure whether the observations, especially the paucity of gene conversion in worm genome, reflect the actual biology and evolutional history or it is just an artifact. However, the abundance of gene conversion in the yeast genome looks real, given the literature pointing out the higher gene conversion rate in yeast (Li 1997, p. 311).

Another remarkable feature is that gene conversion definitely prefers linked gene pairs to unlinked ones in any of the organisms examined. This bias is consistent with the experiments on yeast (Petes and Hill 1988; Haber et al. 1991; Goldman and Lichten 1996). We further observed the negative correlation between the prevalence of gene conversion and the physical distances between duplicated genes in mouse. Similar correlations, with prevalence replaced by frequency, were also reported for the worm genome (Semple and Wolfe 1999) as well as in an experiment on the meiotic recombination in yeast (Goldman and Lichten 1996). These observations are summarized in a statement that a pair of duplicated genes becomes more prone to gene conversion as they get physically closer to each other.

As for the prevalence of gene conversion in vertebrate genomes, Shields (2000) conducted a small-scale analysis before the genome sequences of vertebrates became available. He sampled 20 sets of homologous genes, each set consists of two pairs of human and rodent orthologous genes that are likely to have duplicated prior to the rodent-primate divergence. Using the VTDST3 program (Sawyer 1989) on nucleotide alignments, he found evidence of gene conversion in 45% (=9/20) of his samples. The prevalence appears very high. It appears interesting to compare his result with ours. We, however, became aware that we could not compare them directly, mainly due to the different nature of the data sets. Shields' gene pairs were duplicated before the rodent-primate divergence, whereas ours were duplicated afterward. This means different time intervals of gene conversion events detectable by the two analyses. So we will just mention that our samples are actually as prone to gene conversion as Shields' if we restrict our attention to physically close categories. For example, 32% (=197/607) of our quartets whose mouse pairs lie within 167 kb were positive for gene conversion. The result looks consistent with Shields' ($P = 0.17$ in the binomial test), considering that he collected only physically adjacent gene pairs.

The present study detected two instances of interchromosomal gene conversion (table 2). The frequency of interchromosomal gene conversion is estimated to be 1% (=2/182), which is twice the false-positive ratio of 0.5% that we expect to suffer. Besides, these two mouse pairs were observed in functional categories that are not particularly prone to gene conversion: one belonged to the Kruppel associated box family (two positive out of 17 pairs including this case) and the other to the fructose-biphosphate aldolase, class I family (one positive out of two pairs including this case).

These observations, along with the fact that both of them are moderately positive, make us suspect that they may be false positives. Or, because the refined set of syntenic quartets has no such instance of interchromosomal gene conversion (Supplementary Table S2, Supplementary Material online), they may have resulted from chromosomal rearrangements subsequent to gene conversions. On the other hand, considering the existing evidence for interchromosomal gene conversion in mammals (Arnheim et al. 1980; Murti, Bumbulis, and Schimenti 1994), they may be authentic. Subtelomeric regions may be involved in the phenomenon because they abound with pseudogenes and gene copies. So we examined the dependence on the distance from the chromosomal end (Supplementary Tables S16*A*,*B*, Supplementary Material online). Among the 27 mouse pairs lying within 3 Mb from the chromosomal end, eight (30%) were positive in gene conversion. Thus, the subtelomeric region does seem prone to gene conversion, although the significance is marginal ($P = 0.039$ in Fisher's exact test) due to the small sample size. Then we found that one of the above two positive pairs had one mouse gene within 1 Mb of the chromosome end. But we cannot say anything conclusive due to the lack of statistical power.

Our analysis also revealed that a pair of genes with the opposite transcriptional directions is almost as susceptible to gene conversion as a pair with the same direction. We did not find any analyses addressing this issue in the genome-wide analysis on either yeast or worm, except a conjecture in the paper on the worm saying otherwise (Semple and Wolfe 1999). Although our result might have surprised some people, it may not be so aberrant at least in mammalian (or vertebrate) genomes. There are actually several instances where gene conversion has occurred frequently between duplicated genes or regions with the opposite directions, such as palindrome arms in the male-specific region of the human Y chromosome (Rozen et al. 2003) and inverted duplicons in the human X chromosome (Bagnall et al. 2005). These studies suggest the existence of some mechanisms enabling gene conversion between duplicated genes with the opposite directions. It would be interesting to examine the dependence on the transcriptional directions in creatures other than mammals.

Biological Implications of the Susceptibility Differences Among Functional Categories?

We classified mouse cDNA pairs into functional categories and examined how susceptible each category is to gene conversion. After subtracting the effects of physical distances, family sizes, and synonymous distances, we were able to select a few candidates of functional categories that may have been under selective pressure in favor of/against gene conversion. Such categories are cadherin, vomeronasal receptor, and immunoglobulin V type. The first one favors gene conversion, while the latter two disfavor it. The question is what kind of selective pressure each category has undergone. Protocadherins, which constitute a subfamily of the cadherin family, have recently been suggested to be under the selection pressure that increases the allelic diversity, in which gene conversion may have played a role (Miki et al. 2005). The categories poor in gene conversion might have resulted from the selective pressure that requires the interspecies and/or interlocus divergence of the sequences while keeping their uniformity within the populations. A recent study suggests that vomeronasal receptor type 1 genes, pheromone receptors in rodents, have evolved under positive Darwinian selection to maintain the ability to discriminate between large and complex pheromonal mixtures (Shi et al. 2005). Thus, our conjecture seems true at least for vomeronasal receptors type I.

In order to see whether these functional categories are indeed under selective pressure or not, however, we have to delve further into each of them, looking for functional evidences or evolutionary hallmarks of such selections. In any case, the data presented in this report should become a basis for further data analyses and theoretical as well as experimental studies to elucidate the mechanisms underlying gene conversion.

Supplementary Material

Supplementary Tables S1 through S16 (contained in a file named Ezawa_supplementary_1.xls), Supplementary File named Ezawa_supplementary_2.doc, and Supplementary Figure S1 (in a file named Ezawa_supplementary_3.pdf) are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was mainly conducted by K.E. N.S. proposed the use of quartets, and S.O. helped initial setting of data analyses. The manuscript was mostly written by K.E., with the help of N.S. and S.O. We are grateful to K. Sumiyama and T. Kitano for discussions on this study. We also greatly appreciate three anonymous reviewers, whose comments definitely helped significantly improve this study. This study was supported by a grant in aid for scientific studies from Ministry of Education, Science, Sport, and Culture, Japan, to N.S. K.E. received a fellowship from the Genome Network Project presided by the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Literature Cited

Abe, K., H. Noguchi, K. Tagawa et al. (12 co-authors). 2004. Contribution of Asian mouse subspecies *Mus musculus molossinus* to genomic constitution of strain C57BL/6J, as defined by BAC-end sequence-SNP analysis. Genome Res. **14**:2439–2447.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

Arnheim, N. 1983. Converted evolution of multigene families. Pp. 38–61 *in* M. Nei and R. K. Koehn, eds. Evolution of genes and proteins. Sinauer Associates, Sunderland, Mass.

Arnheim, N., M. Krystal, R. Schmickel, G. Wilson, O. Ryder, and E. Zimmer. 1980. Molecular evidence for genetic exchange among ribosomal genes on nonhomologous chromosomes in man and apes. Proc. Natl. Acad. Sci. USA **77**:7323–7327.

Bagnall, R. D., K. L. Ayres, P. M. Green, and F. Giannelli. 2005. Gene conversion and evolution of Xq28 duplicons involved in recurring inversions causing severe hemophilia A. Genome Res. **15**:214–223.

Balding, D. J., R. A. Nicholas, and D. M. Hunt. 1992. Detecting gene conversion: primate visual pigment genes. Proc. R. Soc. Lond. B Biol. Sci. **249**:275–280.

Birney, E., T. D. Andrews, P. Bevan et al. (48 co-authors). 2004. An overview of Ensembl. Genome Res. **14**:925–928.

Cabrillac, D., V. Delorme, J. Garin, V. Ruffio-Chable, J. L. Giranton, C. Dumas, T. Gaude, and J. M. Cock. 1999. The S1 self-incompatibility haplotype in *Brassica oleracea* includes three S gene family members expressed in stigmas. Plant Cell **11**:971–986.

Charlesworth, D., C. Bartolome, M. H. Schierup, and B. K. Mable. 2003. Haplotype structure of the stigma self-incompatibility gene in natural populations of *Arabidopsis lyrata*. Mol. Biol. Evol. **20**:1741–1753.

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

Doolittle, R. F. 1986. Of URFs and ORFs: a primer on how to analyze derived amino acid sequences. University Science Books, Mill Valley, Calif.

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

Duret, L., D. Mouchiroud, and M. Guoy. 1994. HOVERGEN, a database of homologous vertebrate genes. Nucleic Acids Res. **22**:2360–2365.

Elliott, B., C. Richardson, J. Winderbaum, J. A. Nickoloff, and M. Jasin. 1998. Gene conversion tracts from double-strand break repair in mammalian cells. Mol. Cell Biol. **18**:93–101.

Force, A., M. Lynch, F. B. Pickett, A. Amores, Y.-l. Yan, and J. Postlethwait. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151**:1531–1545.

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

The Gene Ontology Consortium. 2000. Gene ontology: tool for the unification of biology. Nat. Genet. **25**:25–29.

Goldman, A. S. H., and M. Lichten. 1996. The efficiency of meiotic recombination between dispersed sequences in Saccharomyces cerevisiae depends upon their chromosomal location. Genetics **144**:43–55.

Haber, J. E. 2000. Lucky breaks: analysis of recombination in *Saccharomyces*. Mutat. Res. **451**:53–69.

Haber, J. E., W.-Y. Leung, R. H. Boris, and M. Lichten. 1991. The frequency of meiotic recombination in yeast is independent of the number and position of homologous donor sequences: implications for chromosome pairing. Proc. Natl. Acad. Sci. USA **88**:1120–1124.

Hedges, S. B. 2002. The origin and evolution of model organisms. Nat. Rev. Genet. **3**:838–849.

Hubbard, T., D. Andrews, M. Caccamo et al. (52 co-authors). 2005. Ensembl 2005. Nucleic Acids Res. **33**(Database issue):D447–D453.

Hubbard, T., D. Barkeer, E. Birney et al. (35 co-authors). 2002. The Ensembl genome database project. Nucleic Acids Res. **30**:38–41.

Ibbotson, R. D., M. Hunt, J. K. Bowmaker, and J. D. Mollon. 1992. Sequence divergence and copy number of the middle- and long-wave photopigment genes in Old World monkeys. Proc. R. Soc. Lond. B Biol. Sci. **247**:145–154.

Ina, Y. 1994. ODEN: a program package for molecular evolutionary analysis and database search of DNA and amino acid sequences. Comput. Appl. Biosci. **10**:11–12.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature **431**:931–945.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein models. Pp. 21–132 *in* H. N. Munro, ed. Mammalian protein metabolism. Academic, New York.

Kawamura, S., N. Saitou, and S. Ueda. 1992. Concerted evolution of the primate immunoglobulin a-gene through gene conversion. J. Biol. Chem. **267**:7359–7367.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

Kitano, T., and N. Saitou. 1999. Evolution of Rh blood group genes have experienced gene conversions and positive selection. J. Mol. Evol. **49**:615–626.

Kuhner, M. K., D. A. Lawlor, P. D. Ennis, and P. Parham. 1991. Gene conversion in the evolution of the human and chimpanzee MHC class I loci. Tissue Antigens **38**:152–164.

Li, W.-H. 1997. Molecular evolution. Sinauer Associates, Sunderland, Mass.

Liskay, R. M., A. Letsou, and J. L. Stachelek. 1987. Homology requirement for efficient gene conversion between duplicated chromosomal sequences in mammalian cells. Genetics **115**: 161–167.

Lukacsovich, T., and A. S. Waldman. 1999. Suppression of intrachromosomal gene conversion in mammalian cells by small degrees of sequence divergence. Genetics **151**:1559–1568.

Lynch, M., and A. Force. 2000. The probability of duplicated gene preservation by subfunctionalization. Genetics **154**: 459–473.

Martinsohn, J. T., A. B. Sousa, L. A. Guethlein, and J. C. Howard. 1999. The gene conversion hypothesis of MHC evolution: a review. Immunogenetics **50**:168–200.

Miki, R., K. Hattori, Y. Taguchi, M. N. Tada, T. Isosaka, Y. Hidaka, T. Hirabayashi, R. Hashimoto, H. Fukuzato, and T. Yagi. 2005. Identification and characterization of coding single-nucleotide polymorphisms within human protocadherin-α and -β gene clusters. Gene **349**:1–14.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420**:520–563.

Murti, J. R., M. Bumbulis, and J. C. Schimenti. 1994. Gene conversion between unlinked sequences in the germline of mice. Genetics **137**:837–843.

Nei, M., and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**:418–426.

Nei, M., and S. Kumar. 2000. Molecular evolution and phylogenetics. Oxford University Press, New York.

Ohno, S. 1970. Evolution by gene duplication. Splinger-Verlag, Berlin, Germany.

Petes, T. D., and C. W. Hill. 1988. Recombination between repeated genes in microorganisms. Annu. Rev. Genet. **22**: 147–168.

Prigoda, N. L., A. Nassuth, and B. K. Mable. 2005. Phenotypic and genotypic expression of self-incompatibility haplotypes in *Arabidopsis lyrata* suggests unique origin of alleles in different dominance classes. Mol. Biol. Evol. **22**:1609–1620.

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

Reisz, R. R., and J. Muller. 2004. Molecular timescales and the fossil record: a paleontological perspective. Trends Genet. **20**: 237–241.

Reusch, T. B. H., H. Schaschi, and K. M. Wegner. 2004. Recent duplication and inter-locus gene conversion in major histocompatibility class II genes in a teleost, the three-spined stickleback. Immunogenetics **56**:427–437.

Richman, A. D., L. G. Herrera, D. Nash, and M. H. Schierup. 2003. Relative roles of mutation and recombination in generating allelic polymorphism at an MHC class II locus in *Peromyscus maniculatus*. Genet. Res. **82**:89–99.

Rost, B. 1999. Twilight zone of protein sequence alignments. Protein Eng. **2**:85–94.

Rozen, S., H. Skaletsky, J. D. Marszalek, P. J. Minx, H. S. Cordum, R. H. Waterston, R. K. Wilson, and D. C. Page. 2003. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. Nature **423**:873–876.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

Sawyer, S. 1989. Statistical tests for detecting gene conversion. Mol. Biol. Evol. **6**:526–538.

Scott, A. F., P. Heath, S. Trusko, S. H. Boyer, W. Prass, M. Goodman, J. Czelusniak, L.-Y. E. Chang, and J. L. Slightom. 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., and K. H. Wolfe. 1999. Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. J. Mol. Evol. **48**:555–564.

Shi, P., J. P. Bielawski, H. Yang, and Y.-P. Zhang. 2005. Adaptive diversification of vomeronasal receptor 1 genes in rodents. J. Mol. Evol. **60**:566–576.

Shields, D. C. 2000. Gene conversion among chemokine receptors. Gene **246**:239–245.

Shyue, S.-K., L. Li, B. H.-J. Chang, and W.-H. Li. 1994. Intronic gene conversion in the evolution of human X-linked color vision genes. Mol. Biol. Evol. **11**:548–551.

Springer, M. S., W. J. Murphy, E. Eizirik, and S. J. O'Brien. 2003. Placental mammal diversification and the Cretaceous-Tertiary boundary. Proc. Natl. Acad. Sci. USA **100**:1056–1061.

Stephens, J. C. 1985. Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. Mol. Biol. Evol. **2**:539–556.

Takahata, N. 1993. Comments on the detection of reciprocal recombination or gene conversion. Immunogenetics **39**: 146–149.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap-penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

Weiss, E. H., A. Mellor, L. Golden, K. Fahrner, E. Simpson, J. Hurst, and R. A. Flavell. 1983. The structure of a mutant H-2 gene suggests that the generation of polymorphism in H-2 genes may occur by gene conversion-like events. Nature **301**:671–674.

Winderickx, J., L. Battisti, Y. Hibiya, A. G. Motulsky, and S. S. Deeb. 1993. Haplotype diversity in the human red and green opsin genes: evidence for frequent sequence exchange in exon 3. Hum. Mol. Genet. **2**:1413–1421.

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

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

Laura Katz, Associate Editor

Accepted January 5, 2006