JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 17, Number 9, 2010

Pp. 1281–1297
DOI: 10.1089/cmb.2010.0103

An Effective Method for Detecting Gene Conversion Events in Whole Genomes

CHIH-HAO HSU,[1] YU ZHANG,[1] ROSS C. HARDISON,[1]
NISC COMPARATIVE SEQUENCING PROGRAM,[2] ERIC D. GREEN,[2] and WEBB MILLER[1]

ABSTRACT

Gene conversion events are often overlooked in analyses of genome evolution. In a conversion event, an interval of DNA sequence (not necessarily containing a gene) overwrites a highly similar sequence. The event creates relationships among genomic intervals that can confound attempts to identify orthologs and to transfer functional annotation between genomes. Here we examine 1,616,329 paralogous pairs of mouse genomic intervals, and detect conversion events in about 7.5% of them. Properties of the putative gene conversions are analyzed, such as the lengths of the paralogous pairs and the spacing between their sources and targets. Our approach is illustrated using conversion events in primate CCL gene clusters. Source code for our program is included in the 3SEQ_2D package, which is freely available at www.bx.psu.edu/miller_lab/.

Key words: algorithms, computational molecular biology, evolution.

1. INTRODUCTION

SEVERAL CLASSES OF EVOLUTIONARY OPERATIONS have sculpted genomes. Nucleotide substitutions have been studied in great detail for years, and much attention is now focused on large-scale events such as insertions, deletions, inversions, and duplications. Frequently overlooked are gene conversion events (Hurles, 2004; Chen et al., 2007), in which one region is copied over the location of a highly similar region; before the operation there are two genomic intervals, say A and B with 95% identity, and afterwards there are two identical copies of A, one in the position formerly occupied by B.

Conversion events need to be accounted for when attempting to understand the evolution of genomes based on identification of orthologous regions in other species. To take a hypothetical example, suppose mouse genes A and B are related by a duplication event that pre-dated the separation of mouse and rat, so that rat also has genes A and B. A conversion event in a mouse ancestor that overwrote some of B with sequence from A could cause all or part of B's amino-acid sequence to be more closely related to the rat A protein than to the rat B protein, even though B's regulatory regions might remain intact. Successful design

[1]Center for Comparative Genomics and Bioinformatics, Pennsylvania State University, University Park, Pennsylvania.
[2]NIH Intramural Sequencing Center (NISC) and Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland.

and interpretation of experiments in rodents to understand gene B might well require knowledge of these evolutionary relationships.

Gene conversion events have been studied in a variety of species, including the following investigations. Drouin (2002) characterized conversions within 192 yeast gene families; Semple and Wolfe (1999) detected conversion events in 7,397 *Caenorhabditis elegans* genes; Ezawa et al. (2006) studied 2,641 gene quartets, each consisting of two pairs of orthologous genes in mouse and rat, and found that 488 (18%) appear to have undergone gene conversion; and Xu et al. (2008) detected 377 gene conversion events within 626 multigene families in the rice genome. However, these studies investigated gene conversion events only between pairs of protein-coding genes, although conversion can occur between any pair of highly similar regions (Chen et al., 2007). Furthermore, these studies only examined a few thousand pairs of genes, while we cover more than one million paralogous pairs of regions, requiring a more efficient method to deal with such a large data set.

Evidence of conversion between genes frequently appears in cases where the conversion involves only part of a duplicated region. For instance, consider the CC chemokine ligand (CCL) gene cluster in primates. A vervet-vervet alignment reveals similarities extending beyond the genes, created by an ancient duplication event pre-dating the radiation of primates; see Figure 1B. To test whether the elevated percent identity in the protein-coding regions can be explained entirely by purifying selection on those regions, we can compare the pattern of sequence conservation between the paralogous vervet regions with that between vervet CCL15 and its ortholog in an appropriately divergent species. Using dusky titi (a New-World monkey), we see that in most of the interval around the CCL15 gene, the vervet sequence is more similar to the dusky titi CCL15 region (thick line) than to the vervet CCL23 region (thin line) as expected, but this is reversed in a large interval containing exons 2 and 3, and also around the region of exon 1; (Fig. 1C). One reasonable inference from this observation is that conversion events overwrote these intervals with the homologous sequences from the CCL23 gene, or vice versa. Indeed, our procedure identifies a conversion event covering an interval that starts somewhat upstream of exon 2 and extends just beyond exon 3. A smaller interval around exon 1 also shows statistically significant evidence of conversion (Fig. 1D).

A number of statistical tests have been proposed for detecting gene conversions. However, most of these are only efficient for small data sets, e.g., individual gene clusters. Boni et al. (2007) nicely summarize the computational methods available for detecting mosaic structure in sequences, and propose a new method that is particularly economical in terms of computer execution time for large data sets. One drawback is that their algorithm requires large amounts of computer memory. However, we show here that this method can be

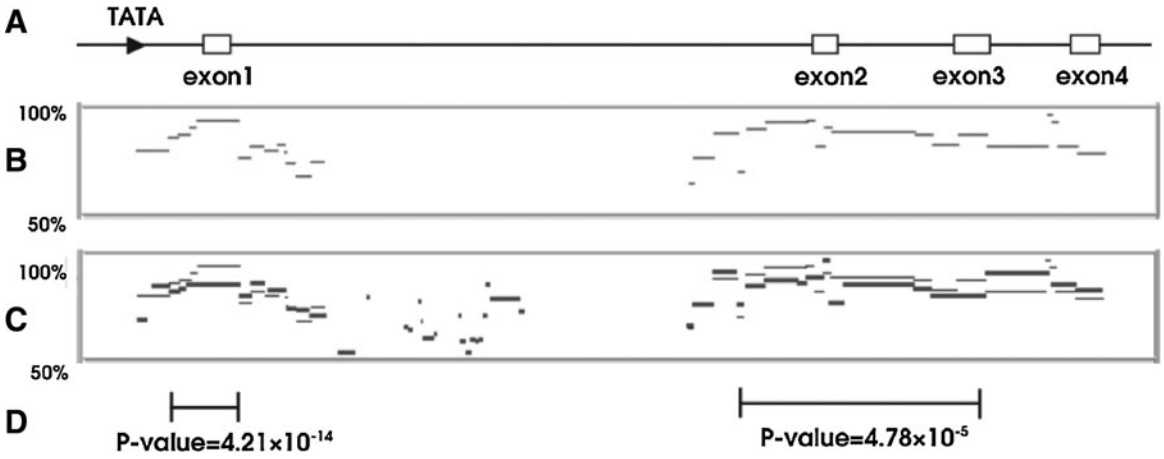

FIG. 1. Evidence of gene conversion in the vervet CCL15 gene. **(A)** Schematic view of the gene. **(B)** Percent identity plot of an alignment to an interval containing the vervet CCL23 gene; each short horizontal line indicates the percent identity over a gap-free subinterval of the alignment. **(C)** Plot of an alignment to the orthologous dusky titi CCL15 gene (thick line) compared to the same paralogous vervet alignment from the previous panel (thin line). **(D)** An interval containing the second and third exons exhibiting gene conversion detected by the method described here, and a smaller interval around the first exon that also shows statistically significant evidence of conversion. See the text for further discussion.

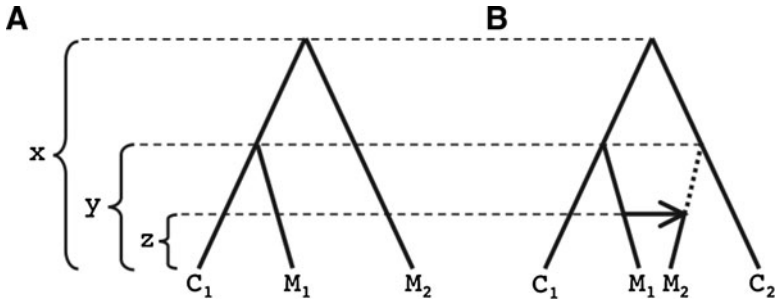

FIG. 2. Timing of evolutionary events. The assumed duplication, speciation, and conversion events between two species, ie, M and C, occurred respectively x, y, and z years ago so that M_1 is orthologous to C_1 and M_2 is orthologous to C_2. See the text for further explanation.

reformulated so that the memory requirements are no longer a limiting factor, which allows us to conduct a comprehensive scan for gene conversion events across the entire mouse genome, starting with 1,616,329 pairs of paralogous mouse intervals. For each pair of paralogous intervals, say M_1 and M_2, we choose a sequence from another species, say C_1, that is believed to be orthologous to M_1. These triplets of sequences are examined to find cases where part of M_1 is more similar to M_2 than to C_1, while another part is more similar to C_1. In such cases, the interval of high $M_1 - M_2$ similarity is inferred to have resulted from a conversion event, as illustrated in Figure 1.

Let us take a closer look at the assumptions implicit in this approach. We are using the sequence C_1 in effect to factor out differences in the rates of evolutionary change along M_1. For instance, some positions of M_1 may be evolving neutrally, but at a rate that depends on the adjacent nucleotides (e.g., hypervariable CpG dinucleotides), while others may be protein-coding or regulatory regions under purifying selection. For a simple model, denote the expected number of substitutions per year at position p by μ_p. Also, suppose that this mutation rate holds for all copies after duplication and speciation events, where the duplication event that separated M_1 and M_2 happened x years ago, while the speciation event separating M_1 and C_1 happened $y < x$ years ago (Fig. 2A). Then the expected number of substitutions between (the descendants of) position p in M_1 and M_2 is $2x \times \mu_p$, while between M_1 and C_1 it is $2y \times \mu_p$. The ratio of those two values is $x/y > 1$, independent of the mutation/fixation rate μ_p. Thus, in the absence of a conversion event, we expect M_1 to differ from M_2 more than from C1 regardless of changes in selective pressure *along the sequence*. However, note that changes in selective pressure *along the tree branches* can produce erroneous signals. For instance, consider a position that is under strong purifying selection, except on the branch from the M_1 / C_1 ancestor to C_1. Then the total path length from M_1 to M_2, as weighted by the branch-specific mutation rates, could be less than that from M_1 to C_1, which our method would incorrectly interpret as evidence of a conversion event. However, these erroneous signals can be reduced by combining the results from two triplets. See Methods for a detailed explanation.

2. METHODS

Boni et al. (2007) developed a time-efficient method for identifying conversions and other recombination events, using the $M_1 - M_2$ and $M_1 - C_1$ alignments to identify "informative" positions in M_1, such that either M_1 and M_2 have one nucleotide while C_1 has another (score -1), or M_1 and C_1 have one nucleotide while M_2 has another (score $+1$). The cumulative sum of these scores along M_1 constitutes what is called a hypergeometric random walk (HGRW [Feller, 1957]) under the assumption that M_1's relationships to M_2 and C_1 are invariant across the interval (Fig. 3). Conversions are detected using the test statistic $x_{m,n,k}$, which is the probability of a *maximum descent* of k occurring by chance for a triplet (M_1, M_2, C_1) with m $+1$s and n -1s. The maximum descent is the maximum decrease of scores across the interval, e.g., the rectangular regions in Figure 3. Boni et al. (2007) give a dynamic-programming algorithm for computing $x_{m,n,k}$, which creates a table that can be consulted for an arbitrary number of triplets.

In order to apply Boni et al.'s method to the entire mouse genome, for each given paralogous pair M_1 and M_2, we needed to find an orthologous sequence C_1 from a species at an appropriate evolutionary distance, i.e., that split from the mouse lineage somewhat after the duplication event and before the conversion. Thus,

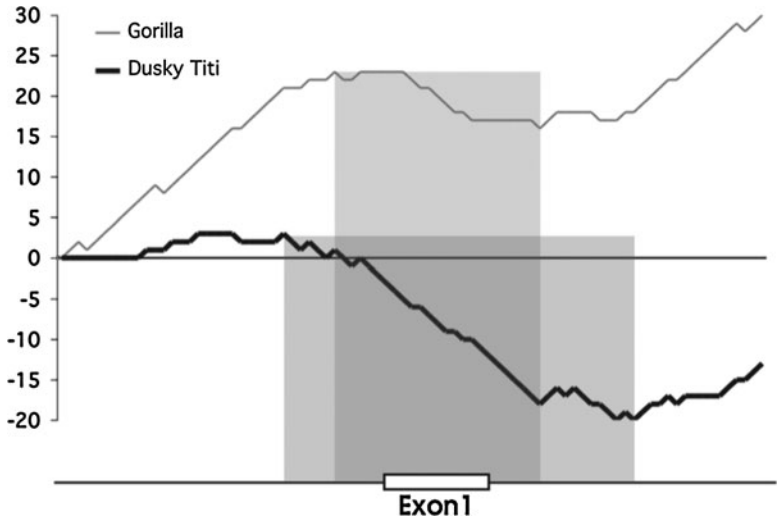

FIG. 3. Maximum descent of the hypergeometric random walks for an alignment between the intervals around the first exon of the vervet CCL15 gene and the vervet CCL23 gene. Using the dusky titi sequence as C_1 identifies a wider converted interval than does using gorilla, possibly because of nested conversion events at different times during the evolution of the vervet lineage.

we tried several available mammalian genome sequences: rat, human, and dog. Each of these species can be used to detect gene conversion events in a particular period of evolution along the lineage leading to mouse. Because the orthologs of M_1 and M_2 often differ, up to 6 triplets were used to look for gene conversion in a given mouse paralogous pair.

2.1. Space-efficient modifications

The original formulation by Boni et al. (2007) requires an amount of computer time and memory that is proportional to B^4, where B is the maximum of m, n, and k. For a triplet with 400 informative sites, this approach would use 6.4 GB of computer memory, allowing the method to work only with relatively short sequences. We modified that method to need only space proportional to $mn + n^2 + SP$ (where $S =$ number of outgroup species and $P =$ number of sequence pairs), as we now describe.

In the notation of Boni et al. (2007), the test statistic $x_{m,n,k}$ is defined as $P(md\ H_{m,n} = k)$, which is the probability for a hypergeometric random walk with m up steps and n down steps, i.e., $H_{m,n}$, to have maximum descent of k, and can be calculated using the equation:

$$x_{m,n,k} = \sum_{j=0}^{k} y_{m,n,k,j} \tag{1}$$

where:

$$y_{m,n,k,j} = P\left(md\ H_{m,n} = k \bigcap min\ H_{m,n} = -j\right) \tag{2}$$

The probabilities y, which places one more constriction on the minimum value of $H_{m,n}$ to be $-j$, can be obtained by dynamic programming based on the following recursive relationships.

$$y_{m,n,k,j} = \begin{cases} \left(\frac{m}{m+n}\right)[y_{m-1,n,k,1} + y_{m-1,n,k,0}] & \text{if } j = 0 \\ \left(\frac{m}{m+n}\right) y_{m-1,n,k,j+1} + \left(\frac{n}{m+n}\right) y_{m,n-1,k,j-1} & \text{if } k > j > 0 \\ \left(\frac{n}{m+n}\right)[y_{m,n-1,k-1,j-1} + y_{m,n-1,k,j-1}] & \text{if } j = k > 0 \\ 0 & \text{if } j > k \geq 0 \end{cases} \tag{3}$$

With boundary conditions

$$y_{m,0,k,j} = \begin{cases} 1 & \text{for } k = j = 0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$y_{0,n,k,j} = \begin{cases} 1 & \text{for } k = j = n \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$y_{m,n,0,0} = \begin{cases} 1 & \text{for } n = 0 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

$$y_{m,n,k,j} = 0 \text{ when } k > n \text{ or } k < n - m. \tag{7}$$

$$y_{m,n,k,j} = 0 \text{ when } j > n \text{ or } j < n - m. \tag{8}$$

Thus, the p-value for a triplet with $m+1$s, $n-1$s, and maximum descent of k is defined as $P(mdH_{m,n} \geq k)$ and can be calculated using the following equation:

$$p - value_{m,n,k} = \sum_{j=k}^{n} x_{m,n,j} \tag{9}$$

In order to reduce memory usage, we introduce the additional variable $A_{m,n,k}$, defined as:

$$A_{m,n,k} = y_{m,n,k,k} = \left(\frac{n}{m+n}\right)[A_{m,n-1,k-1} + y_{m,n-1,k,k-1}] \tag{10}$$

Then,

$$x_{m,n,k} = \sum_{j=0}^{k} y_{m,n,k,j} = \left(\frac{m}{m+n}\right)x_{m-1,n,k} + \left(\frac{n}{m+n}\right)[x_{m,n-1,k} - A_{m,n-1,k} + A_{m,n-1,k-1}] \tag{11}$$

The key observation is that, for fixed k, the only component of the equation 3 that depends on $k-1$ is when $j = k > 0$, and in that case the required value is $A_{m,n-1,k-1}$. Consequently, provided that we record the 3-dimensional array of values $A_{m,n,k}$, we can store the values of y for a fixed k in another 3-dimensional array that we call $y_{m,n,j}$ and overwrite them with the values corresponding to $k+1$ as the computation proceeds. The resulting algorithm, given in Figure 4, uses only two arrays of size mn^2 (x and y can be stored in the same array). It can handle triplets with 2000 informative sites on a mid-sized workstation.

Furthermore, since the value of x depends only on the values in the same loop, i.e., $x_{m,n-1,k}$, and in the previous loop, i.e., $x_{m-1,n,k}$ (when using m as the outer loop), an $O(mn + n^2 + SP)$ space method (where $S =$ number of outgroup species and $P =$ number of paralogous pairs) is possible. Instead of using a three dimensional array to store all values of $x_{m,n,k}$, the combinations of (m, n, k) that actually occur in the dataset of triplets (M_1, M_2, C_1) are determined and stored as nodes in a three-dimensional linked list data structure, as shown in Figure 5 (triplets with the same values of (m, n, k) are grouped in the same node, and all nodes are linked in ascending order in three dimensions). This consumes $O(SP)$ space. To obtain the p-values for all nodes, i.e., equation 9, the values of $x_{m,n,k}$ are calculated and summed to the relevant nodes, i.e., the nodes between $(m, n, 0)$ and (m, n, k). For this purpose, a two dimensional array called *Linked_List_Table*, which points to the starting position for each pair of (m, n), is maintained so that the relevant nodes for a particular $x_{m,n,k}$ can be retrieved quickly. A detailed algorithm is shown in Figure 6. Since only those values necessary for further calculation are kept (the values of $x_{m-1,n,k}$ and $x_{m,n,k}$ are stored in the two-dimensional arrays of $x_{0,n,k}$ and $x_{1,n,k}$, respectively; $x_{0,n,k}$ and $x_{1,n,k}$ are overwritten with the values corresponding to m and $m+1$ as the computation proceeds), the maximum table size required for the calculation of x is $O(mn + n^2)$.

Although the space requirement is thus reduced, the time complexity is still quartic (exponent 4). Also, the longest interval in our data is 394,252 base pairs. In order to deal with long alignments, those with length greater than 5000 are divided into several sub-alignments with 1000 sites overlapped. The p-value for each sub-alignment is then calculated, and a multiple-comparison correction method (Holm, 1979) is used to determine if the set of sub-alignments supports an assertion that the whole alignment shows significant signs of a conversion.

2.2. *Extension to quadruplet testing*

It is not uncommon that we have a pair of paralogs in the other species, say C_1 and C_2 in rat, that are orthologs for M_1 and M_2 in mouse, respectively. In a fashion similar to the triplet testing, we can perform

```
Cubic-HGRW (MAX_M, MAX_N)
 1    for m ← 0 to MAX_M do
 2        A[m, 0, 0] ← 1
 3        for n ← 1 to MAX_N do
 4            A[m, n, 0] ← 0
 5    for k ← 1 to MAX_N do
 6        for m ← 0 to MAX_M do
 7            for j ← 0 to MAX_N do
 8                y[m, 0, j] ← 0
 9        for n ← 0 to MAX_N do
10            for j ← 0 to MAX_N do
11                if k = n and j = n then
12                    y[0, n, j] ← 1
13                else    y[0, n, j] ← 0
14        for m ← 1 to MAX_M do
15            for n ← 1 to MAX_N do
16                for j ← 0 to MAX_N do
17                    if k > n or k < n – m then
18                        y[m, n, j] ← 0
19                    else if j > k or j > n or j < n - m then
20                        y[m, n, j] ← 0
21                    else if j = 0 then
22                        y[m, n, j] ← m / (m + n) × (y[m - 1, n, 1] + y[m - 1, n, 0])
23                    else    y[m, n, j] ← m / (m + n) × y[m - 1, n, j + 1]
                                           + n / (m + n) × y[m, n - 1, j - 1]
24                if k > n or k < n – m then
25                    A[m, n, k] ← 0
26                else    A[m, n, k] ← n / (m + n) × (A[m, n - 1, k - 1] + y[m, n - 1, k - 1])
27                y[m, n, k] ← A[m, n, k]
28    for n ← 0 to MAX_N do
29        for k ← 0 to MAX_N do
30            if n = k then
31                x[0, n, k] ← 1
32            else    x[0, n, k] ← 0
33    for m ← 1 to MAX_M  do
34        x[m, 0, 0] ← 1
35        for k ← 1 to MAX_N do
36            x[m, 0, k] ← 0
37        for n ← 1 to MAX_N do
38            x[m, n, 0] ← 0
39            for k ← 1 to MAX_N do
40                x[m, n, k] ← m / (m + n) × x[m - 1, n, k] + n / (m + n)
                               × (x[m, n - 1, k] + A[m, n - 1, k - 1] - A[m, n - 1, k])
41    return x
```

FIG. 4. A cubic-space algorithm for computing the probabilities $x_{m,n,k}$.

quadruplet testing (M_1, M_2, C_1, C_2) that is the summation of the hypergeometric random walks of two triplets, i.e., (M_1, M_2, C_1) and (M_1, M_2, C_2), as shown in Figure 7. This joint testing of gene conversions often has higher power than triplet testing. For example, in a four-way alignment if we observe that $M_{1i} = M_{2i} \neq C_{1i} = C_{2i}$ at a column i, this is strong evidence for gene conversion.

We again need to assign a score to each column in the four-way alignment in order to calculate maximum descent scores. We use the sum of the triplet scores. In the quadruplet case of (M_1, M_2, C_1, C_2), there are two distinct triplets of interest: (M_1, M_2, C_1) and (M_1, M_2, C_2). Other possible triplets are ignored, as we are testing for gene conversions in mouse. In the case of $M_{1i} = M_{2i} \neq C_{1i} = C_{2i}$, we assign a score of -2 to column i, because each triplet ($M_{1i} = M_{2i} \neq C_{1i}$ and $M_{1i} = M_{2i} \neq C_{2i}$) has score -1. On the other hand, if $M_{1i} = C_{1i} \neq M_{2i} = C_{2i}$, we assign a score of $+2$ by the same rule. Furthermore, if $M_{1i} \neq M_{2i} = C_{1i} = C_{2i}$ or $M_{2i} \neq M_{1i} = C_{1i} = C_{2i}$, we assign score $+1$ because one triplet has score $+1$ and the other has score 0 (non-informative). In summary, all columns are assigned the sum of their two triplet scores, and all columns with 0 score are subsequently ignored.

Suppose that there are m_1 one-step-ups, m_2 two-step-ups, n_1 one-step-downs, and n_2 two-step-downs for a particular quadruplet testing. Let $N = m_1 + m_2 + n_1 + n_2$ denote the total number of moves. We calculate the exact p-value of observing at least k maximum descent in a random walk, constrained by (m_1, m_2, n_1, n_2) moves of each type, using the following recursive formula:

$$x_{m_1,m_1,n_1,n_2,k} = \sum_{j=0}^{k} y_{m_1,m_1,n_1,n_2,k,j} \tag{12}$$

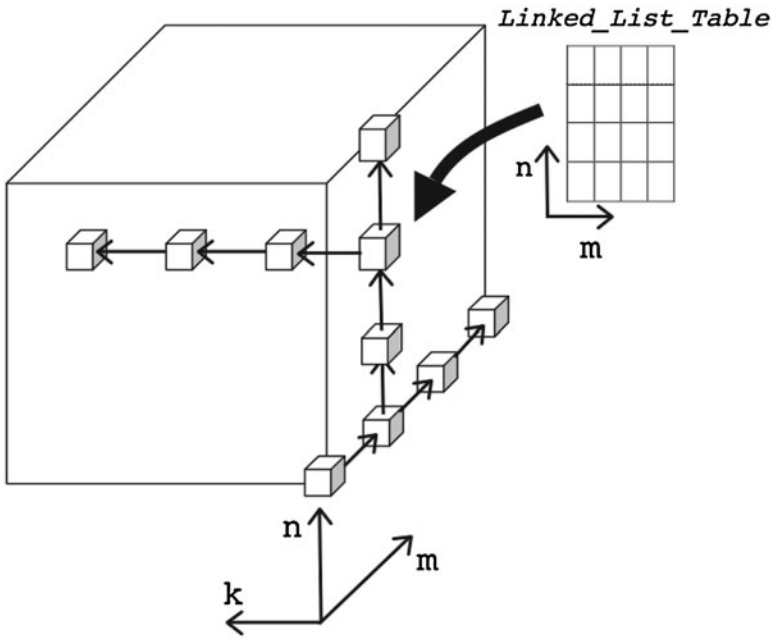

FIG. 5. A 3-D linked list data structure to store necessary values for computing the probabilities $x_{m,n,k}$.

```
Quadratic-HGRW (MAX_M, MAX_N)
 1    for m ← 0 to MAX_M do
 2        A[m, 0, 0] ← 1
 3        for n ← 1 to MAX_N do
 4            A[m, n, 0] ← 0
 5    for n ← 0 to MAX_N do
 6        for k ← 0 to MAX_N do
 7            if n = k then
 8                x[0, n, k] ← 1
 9            else    x[0, n, k] ← 0
10    for k ← 1 to MAX_N do
11        for n ← 0 to MAX_N do
12            for j ← 0 to MAX_N do
13                if k = n and j = n then
14                    y[0, n, j] ← 1
15                else    y[0, n, j] ← 0
16        for m ← 1 to MAX_M do
17            for j ← 0 to MAX_N do
18                y[1, 0, j] ← 0
19            x[1, 0, 0] ← 1
20            x[1, 0, k] ← 0
21            for n ← 1 to MAX_N do
22                for j ← 0 to MAX_N do
23                    if k > n or k < n – m then
24                        y[1, n, j] ← 0
25                    else if j > k or j > n or j < n - m then
26                        y[1, n, j] ← 0
27                    else if j = 0 then
28                        y[1, n, j] ← m / (m + n) × (y[0, n, 1] + y[0, n, 0])
29                    else    y[1, n, j] ← m / (m + n) × y[0, n, j + 1]
                                  + n / (m + n) × y[1, n - 1, j - 1]
30                if k > n or k < n – m then
31                    A[m, n, 1] ← 0
32                else    A[m, n, 1] ← n / (m + n) × (A[m, n - 1, 0] + y[1, n - 1, k - 1])
33                y[1, n, k] ← A[m, n, 1]
34                x[1, n, k] ← m / (m + n) × x[0, n, k] + n / (m + n) × (x[1, n - 1, k] + A[m, n - 1, 0]
                                  - A[m, n - 1, 1])
                  // The value of x[1, n, k] is summed to the nodes between (m, n, 0) and (m, n, k)
35                Update_Linked_List(m, n, k, x[1, n, k])
                // Throw away unnecessary information
36              for n ← 0 to MAX_N do
37                  A[m, n, 0] ← A[m, n, 1]
38                  x[0, n, k] ← x[1, n, k]
39                  for j ← 0 to MAX_N do
40                      y[0, n, j] ← y[1, n, j]
41    return x
```

FIG. 6. A quadratic-space algorithm for computing the probabilities $x_{m,n,k}$.

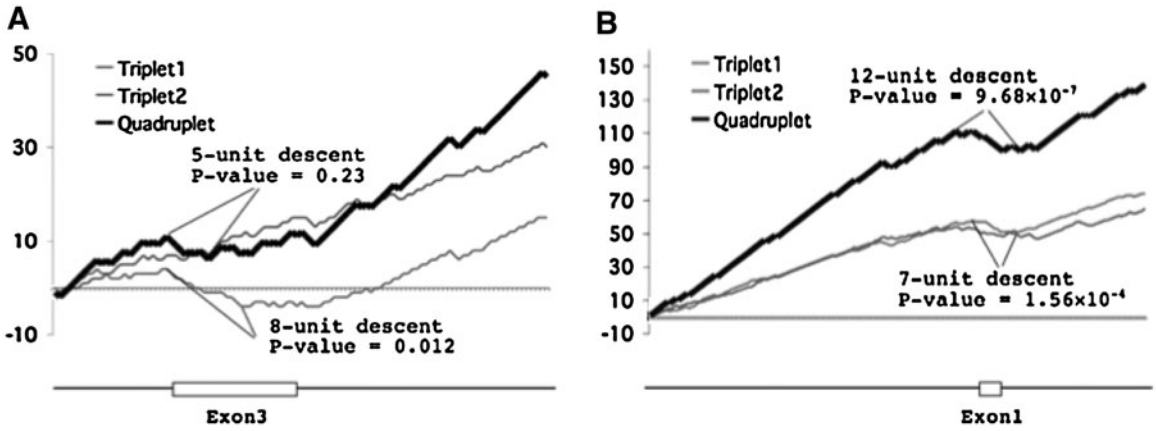

FIG. 7. Comparisons between quadruplet testing and triplet testing. **(A)** The colobus monkey CCL3 and a partial pseudo gene paralog pair. **(B)** The vervet CCL15 and vervet CCL23 paralog pair.

where

$$y_{m_1,m_1,n_1,n_2,k,j} = \begin{cases} \frac{m_1}{N}\sum_{l=0}^{1} y_{m_1-1,m_2,n_1,n_2,k,l} + \frac{m_2}{N}\sum_{l=0}^{2} y_{m_1,m_2-1,n_1,n_2,k,l} & \text{if } j=0 \\ \frac{m_1}{N} y_{m_1-1,m_2,n_1,n_2,k,j+1} + \frac{m_2}{N} y_{m_1,m_2-1,n_1,n_2,k,j+2} + & \\ \frac{n_1}{N} y_{m_1,m_2,n_1-1,n_2,k,j-1} + \frac{n_2}{N} y_{m_1,m_2,n_1,n_2-1,k,j-2} & \text{if } k>j>0 \\ \frac{n_1}{N}\sum_{l=k-1}^{k} y_{m_1,m_2,n_1-1,n_2,l,j-1} + \frac{n_2}{N}\sum_{l=k-2}^{k} y_{m_1,m_2,n_1,n_2-1,l,j-2} & \text{if } j=k>0 \\ 0 & \text{if } j>k\geq 0 \end{cases} \quad (13)$$

The y terms are set to 0 if their subscripts go below 0 or above k. Let $n = n_1+2n_2$ and $m = m_1+2m_2$, then k is bounded between $[max(n-m, 1), n]$ for $m>0$, $n>0$, and the computation time for the p-values of all possible (m_1, m_2, n_1, n_2, k) combinations is $O(m_1m_2n_1n_2n^2)$, and the memory usage is $O(m_1m_2n_1n_2n)$. However, since the time complexity and memory consumption for this formula are very high in practice, we use the same formula as in triplet testing, i.e., equation 11, to get p-values in our program, even though it is more conservative.

Quadruplet testing often has higher specificity and sensitivity than triplet testing for detecting conversions. For example, in Figure 7A, a weakly significant (0.012) conversion event in colobus monkey was detected between CCL3 and a partial pseudo gene in one triplet, but there is no evidence for the event in the other triplet. This could be due to a faster evolutionary rate in the (probably non-functional) partial gene than in the coding region of CCL3. Quadruplet testing did not show any evidence of conversion in this region, which suggests that the effect of one triplet can be neutralized by that of the other triplet when there is no conversion between a paralog pair. On the other hand, in cases where the triplets reinforce each other, quadruplet testing can give a more significant result, as shown in Figure 7B. Therefore, whenever orthologs for both M_1 and M_2 are available in a particular outgroup species, we combine the results of the two triplets to perform quadruplet testing.

2.3. Multiple-comparison correction

When several statistical tests are performed simultaneously, a multiple-comparison correction should be applied. In our study, three outgroup species are used, and for this multiplicity we use the Bonferroni correction (Holm, 1979); we multiply the smallest p-value for each paralogous mouse pair by the number of tests (up to 3), and use this adjusted p-value to evaluate the significance of any potential gene conversion in that pair.

Multiple-comparison correction is also applied to compensate for the many pairs of paralogous sequences. For the 1,616,329 pairs that were analyzed, we used a correction method that controls the false discovery rate (FDR), proposed by Benjamini and Hochberg (1995). The cutoff threshold for p-values can be found by the following algorithm:

Algorithm. CutOff(α, p-values)

```
1 sort p-values in ascending order
2 for i ← 1 to number of p-values do
3       if p_i > (i ÷ number of p-values)×α
4    return (i ÷ number of p-values)×α
```

In our mouse study, α was set to 0.05 and the corresponding cutoff threshold for p-values was 0.003771. This means that only a test whose p-value after Bonferroni correction was less than 0.003771 was considered as significant evidence for gene conversion.

2.4. Directionality of gene conversion

We attempt to determine the source and target of a conversion event as follows. As shown in Figure 2B, let us suppose that duplication, speciation, and conversion events occurred x, y, and z years ago respectively, with $x > y > z$, and consider a converted position. Regardless of the direction of the conversion (from M_1 to M_2, or vice versa), M_1 and M_2 are separated by $2z$ total years in the converted region. If M_1 converted M_2 (i.e., part of M_1 overwrote part of M_2), then the separation of M_1 and C_1 is $2y$ but the separation of M_2 and its ortholog, C_2, is $2x > 2y$. This observation serves as a basis for determining the conversion direction. Figure 8 shows an example of determining the source and target of a conversion from CCL23 to CCL15 in vervet.

Specifically, assume (m_1, n_1) with maximum descent k_1 in the first triplet (M_1, M_2, C_1), and (m_2, n_2) with maximum descent k_2 in the second triplet (M_1, M_2, C_2). Note that m_i and n_i here are not the m and n in equation 1; rather, they are the numbers of ups and downs within the common maximum descent regions of the two triplets (intersection). The probabilities of going down in these regions are:

$$p_1 = n_1 \div (m_1 + n_1) \tag{14}$$

$$p_2 = n_2 \div (m_2 + n_2) \tag{15}$$

When combining these data in a quadruplet, there are a total of $(m_1 + m_2)$ ups and $(n_1 + n_2)$ downs, and the possibility of going down in the combined data is:

$$p = (n_1 + n_2) \div (m_1 + m_2 + n_1 + n_2) \tag{16}$$

As shown in Figure 2B, if M_1 converted M_2, the separation of M_1 and C_1 is closer than the separation of M_2 and C_2 in the converted region. Thus, p_1 should be smaller than p_2. Our objective function (O) is therefore to determine how significant the difference of $p_1 - p_2$ is, based on the binomial distribution:

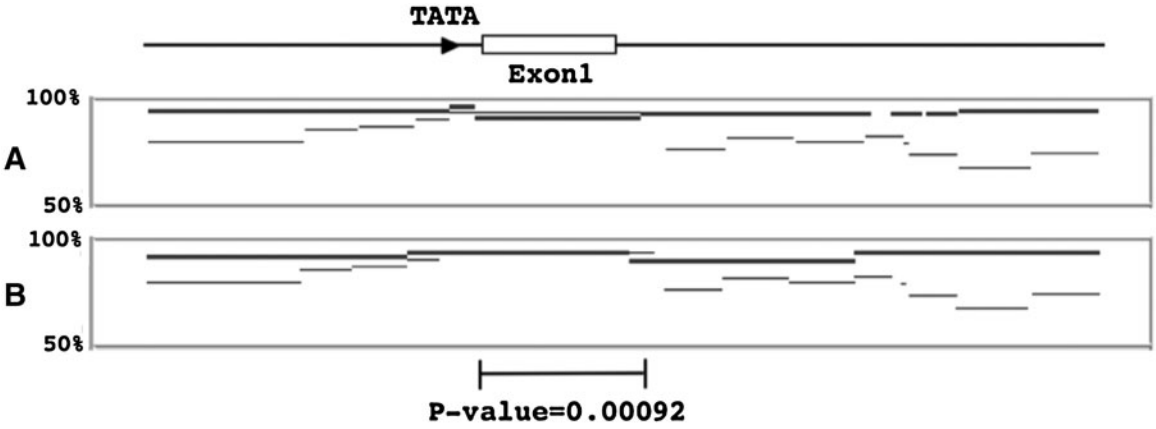

FIG. 8. Evidence that the vervet CCL23 gene converted the vervet CCL15 gene. Percent identity plots for **(A)** CCL15 and **(B)** CCL23 showing alignments to the vervet paralog as a thin line, and alignments to the putative colobus ortholog as a thick line. In the converted region, the vervet-colobus alignments have 94% identity for CCL15 and 97% identity for CCL23.

$$O = ((p_1 - p_2) - E(p'_1 - p'_2)) \div sqrt(V(p'_1 - p'_2)) \tag{17}$$

where

$$E(p'_1 - p'_2) = 0 \tag{18}$$

$$V(p'_1 - p'_2) = \left(\frac{1}{m_1 + n_1} + \frac{1}{m_2 + n_2}\right) \times p \times (1 - p) \tag{19}$$

In our study, three outgroup species are used to detect gene conversions in the mouse genome. We use the species that shows the most significant difference of $p_1 - p_2$ to determine the directionality of conversion for a given paralogous pair. However, there are several reasons why the direction of a conversion might not be clear, even when using several outgroup species, such as conversions in the outgroup species or missing outgroup data. Only part of the direction for the putative conversions can be determined.

3. RESULTS

3.1. Highly conserved pairs of sequences

We aligned each pair of mouse chromosomes that are masked out with REPEATMASKER (Smit, 1999), including self-alignments, using BLASTZ (Schwartz et al., 2003) with T = 2 and default values for the other parameters. Alignments with identity of less than 70% were removed. Chaining of the mouse-mouse alignments was performed using the method of Zhang et al. (1994). For alignments between mouse intervals and their putative orthologs in other species, we used the pairwise alignment nets (Kent et al., 2003) downloaded from the UCSC Genome Browser website (Kent et al., 2002). We applied the modified method described in this paper and recorded information about the inferred conversion events in Table 1.

Of the 1,616,329 analyzed pairs of mouse sequences, 121,899 (7.5%) indicated a gene conversion event. The fraction of intra-chromosomal pairs indicating a conversion (13.8%) is significantly higher than for inter-chromosomal pairs (6.7%).

3.2. Association with gene conversion

To study the correlations between various genomic features and gene conversion, we used logistic regression models (Agresti, 2002) to characterize gene conversions based on the following factors.

strand:	binary; strand of the second paralog relative to the first one
seq_len:	paralog size in basepairs
pair_dist:	distance between the paralogs
seq_sim:	percent identity between the two paralogs
gc:	percentage of G + C in both paralogs combined
*gc*1:	percentage of G + C in the first paralog
*gc*2:	percentage of G + C in the second paralog
*coding*1:	binary; whether or not the first paralog contains coding regions
*coding*2:	binary; whether or not the second paralog contains coding regions

To make the analysis robust, we first binned the continuous factors into ordered categories shown as Table 2.

For ease of interpretation, we only included the main effects of the variables, and carried out the analysis separately for inter-chromosome and intra-chromosome pairs, as shown below under (1) and (2), respectively. For a logistic regression model, the response is binary. In our case, it is gene conversion (indicated

Table 1. Distribution of Intra- and Inter-Chromosomal Gene Conversions

	Intra-chromosome	*Inter-chromosome*	*Total*
Gene conversion	25,189 (13.8%)	96,710 (6.7%)	121,899 (7.5%)
No gene conversion	157,833 (86.2%)	1,336,597 (93.3%)	1,494,430 (92.5%)
Total	183,022	1,433,307	1,616,329

TABLE 2. CATEGORIES FOR CONTINUOUS PARALOG PROPERTIES

Categories	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*
seq_len	<200	200–500	500-1k	1k–2k	2k–5k	≥5k		
pair_dist	inter	<1k	1k–10k	10k–100k	100k–1m	1m–10m	10m–100m	≥100m
seq_sim	<0.75	0.75–0.8	0.8–0.85	0.85–0.9	0.9–0.95	0.95–1		
gc, *gc*1, *gc*2	<0.4	0.4–0.45	0.45–0.5	0.5–0.55	0.55–0.6	≥0.6		

by 1) or not (indicated by 0). The logistic regression relates the logit of the probability of gene conversion to a function of predictors, where the logit function is $logit(x) = log(x \div (1 - x))$. The actual gene conversion event is regarded as a binary outcome with the probability given by the regression model.

(1) For inter-chromosome paralog pairs, the model is:

$$logit(conversion\_rate) \sim seq\_len + seq\_sim + gc + coding1 + coding2$$

Coefficients:

	Estimate	*Std. error*	*z Value*	*Pr(>|z|)*
(Intercept)	−3.598053	0.008202	−438.668	< 2e-16***
seq_len	0.730605	0.003035	240.759	< 2e-16***
seq_sim	0.114288	0.002934	38.959	< 2e-16***
gc	0.050462	0.003065	16.462	<2e-16***
*coding*1	0.122936	0.021215	5.795	6.84e-09***
*coding*2	0.361988	0.020886	17.332	< 2e-16***

If the estimated coefficient of a variable is positive, the variable increases the gene conversion probability (or rate), while a negative value indicates a decrease.

(2) For intra-chromosome paralog pairs, we included two more variables, *strand* and *pair_dist*, the model is:

$$logit(conversion\_rate) \sim strand + seq\_len + seq\_sim + gc + coding1 + coding2 + pair\_dist$$

Coefficients:

	Estimate	*Std. error*	*z Value*	*Pr(>|z|)*
(Intercept)	−1.719952	0.030422	−56.536	< 2e-16***
strand	0.032592	0.014835	2.197	0.028*
seq_len	0.520803	0.004999	104.172	< 2e-16***
seq_sim	0.030534	0.006159	4.958	7.12e-07***
gc	0.158865	0.006520	24.365	< 2e-16***
*coding*1	0.277185	0.023922	11.587	< 2e-16***
*coding*2	0.193312	0.024272	7.964	1.66e-15***
pair_dist	−0.269120	0.005643	−47.688	< 2e-16***

Based on the results of (1) and (2), we see that:

- The conversion rate is higher when M_1 and M_2 are on the same chromosome; this can be seen from the larger intercept of (2) than of (1).
- Strand has little effect on gene conversion. It seems natural that relative strand is not a factor when the paralogs are on different chromosomes. When M_1 and M_2 are on the same chromosome, from (2) we see that strand effect is positive (0.028) but only weakly significant.
- Conversion rate increases as paralog size increases; this can be seen from the positive coefficient of *seq_len* in both (1) and (2).
- Similarity of sequences has a significant effect on conversion rate for both inter- and intra-chromosome pairs.

- Both GC content and the presence of coding sequences contribute positively to the conversion rate for inter- and intra-chromosome pairs.
- Conversion rate decreases as the distance between two paralogs increases; this can be seen from (2) where the coefficient of *pair_dist* is negative.

The models used here are simplified; they do not account for interactions among factors. We did not include interactions because (a) they are more complicated to interpret, and (b) they would require much more computer memory, considering that over a million paralog pairs were being tested. Instead, we performed small-scale studies using subsets of the data, and we observed that although some interactions are significantly related to gene conversion rate, the magnitude of their contributions is relatively small compared to the factors direct effects. Furthermore, we did not account for differences among chromosomes, although conversion rates do vary significantly depending on the chromosome.

3.3. Directionality of gene conversion

To obtain a logistic regression model for the conversion direction (in the cases where it could be determined), we used the discrete variable *con_direction*, set to 1 if H_2 converts H_1, and 0 if H_1 converts H_2.

(1) For inter-chromosome paralog pairs, the model is:

$$logit(con\_direction) \sim seq\_len + seq\_sim + gc1 + gc2 + coding1 + coding2$$

Coefficients:

	Estimate	*Std. error*	*z Value*	*Pr(>|z|)*
(Intercept)	0.0098630	0.0178540	0.552	0.5807
seq_len	−0.0001498	0.0074786	−0.020	0.9840
seq_sim	−0.0120501	0.0062416	−1.931	0.0535
*gc*1	−0.6115270	0.0109130	−56.037	< 2e-16***
*gc*2	0.6031434	0.0108527	55.576	< 2e-16***
*coding*1	−1.1629914	0.0487090	−23.876	< 2e-16***
*coding*2	1.2867614	0.0481985	26.697	< 2e-16***

(2) For intra-chromosome paralog pairs, the model is:

$$logit(conversion\_rate) \sim strand + seq\_len + seq\_sim + gc1 + gc2 + coding1 + coding2 + pair\_dist$$

Coefficients:

	Estimate	*Std. error*	*z Value*	*Pr(>|z|)*
(Intercept)	0.213053	0.072980	2.919	0.00351**
strand	−0.096799	0.031290	−3.094	0.00198**
seq_len	−0.017195	0.011786	−1.459	0.14461
seq_sim	−0.003534	0.013072	−0.270	0.78691
*gc*1	−0.604822	0.028992	−20.862	< 2e-16***
*gc*2	0.612293	0.028947	21.152	< 2e-16***
*coding*1	−0.568145	0.049264	−11.533	< 2e-16***
*coding*2	0.398675	0.049120	8.116	4.8e-16***
pair_dist	−0.029474	0.013091	−2.251	0.02435*

Based on the results of (1) and (2), we see that:

- Sequence length and similarity are not significantly associated with conversion direction.
- Strand and pair distance have slight effects on conversion direction.
- GC content and coding sequences significantly affect the direction for both inter- and intra-chromosome pairs. Negative *gc*1 and positive *gc*2 suggest that the conversion direction is more likely to be from the region with higher GC-content to the one with lower GC content. Negative *coding*1 and positive *coding*2 mean that the conversion direction tends to be from a functional gene to a pseudogene.

3.4. Analysis of CCL gene cluster (hg18.chr17:31,334,806-31,886,998)

In order to evaluate the performance of our method, we tested it on the CCL gene cluster in seven primates. We searched for conversion events in each of the species, using the other six as outgroups. We also compared our results to those from another gene conversion detection method, GARD (Pond et al., 2006), for confirmation.

The CCL gene cluster contains several chemokine ligand genes, some of which affect the ability of the HIV virus to enter the cell (Modi et al., 2006). In our analysis, we found that this gene cluster expanded after the separation of humans and Old World monkeys, and detected a number of conversion events shown in Figure 9. Within the gene-cluster sequences of the seven primates, we identified 42 genes and 5 pseudogenes using GeneWise2 (Birney et al., 2004). Our results show that several conversion events occurred between CCL15 and CCL23 in different lineages. In addition, a number of gene conversion events were detected in non-coding regions.

Based on our results, we found that several conversion events occurred involving the first exon in the CCL15 and CCL23 genes along various species. Also, an earlier conversion event occurred in the region containing the second and third exons followed by a small conversion event around the second exon; there is no conversion in the fourth exon. To show the correctness of our results, we separated the alignment of CCL15 and CCL23 genes into four regions as shown in Figure 10A and constructed a phylogenetic tree for each region using the PHYLIP package (Felsenstein, 1989). The phylogenetic trees are shown in Figure 10B–E and they are different from each other. Also, we used the GARD program to confirm this result. The GARD program reports evidence of three breakpoints, shown in Figure 11; they almost separate the whole alignment into four regions near the boundaries of exons. Both of these two results indicate that the evolutionary histories of these four regions are different. Our results can give a reasonable explanation about the inconsistencies among these four phylogenetic trees. The CCL15 and CCL23 genes were separated by an ancient duplication event pre-dating the radiation of primates. Several gene conversion events occurred in the first three exons after the split of simian primates and lemurs. Following the separation of New World and Old World monkeys, a conversion event occurred in Old World monkeys covering an interval that starts somewhat upstream of exon 2 and extends just beyond exon 3, and another one occurred in the region containing exon 1. Finally, after the separation of apes and Old World monkeys, there was a conversion event in the region containing exon 2 in the human lineage, and another in the vicinity of exon 1 in the Old World monkeys. The inferred evolutionary history of the primate CCL15 and CCL23 is shown in Figure 12.

Therefore, combining this with our previous studies, a possible evolutionary history for the entire CCL gene cluster is shown in Figure 13. There were six genes in the root node (primates), and the cluster has undergone significant expansion after the separation of human and Old World monkeys. Several gene conversion events occurred along various lineages.

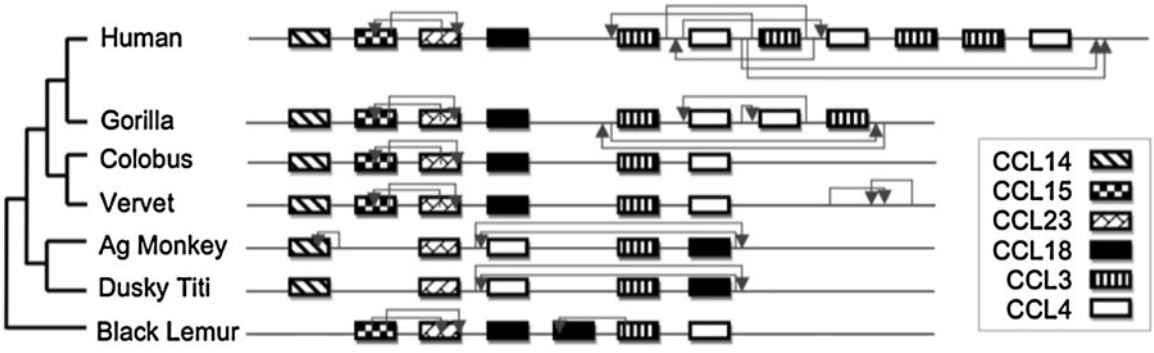

FIG. 9. Gene conversion events detected in the primate CCL gene cluster. Arrows show the directionality of conversion. Several gene conversion events occurred between the coding sequences of the CCL15 and CCL23 genes. Also, a number of conversion events occurred in the non-coding regions or between the flanking non-coding sequence and intron, i.e., around the CCL14 gene of Ag Monkey.

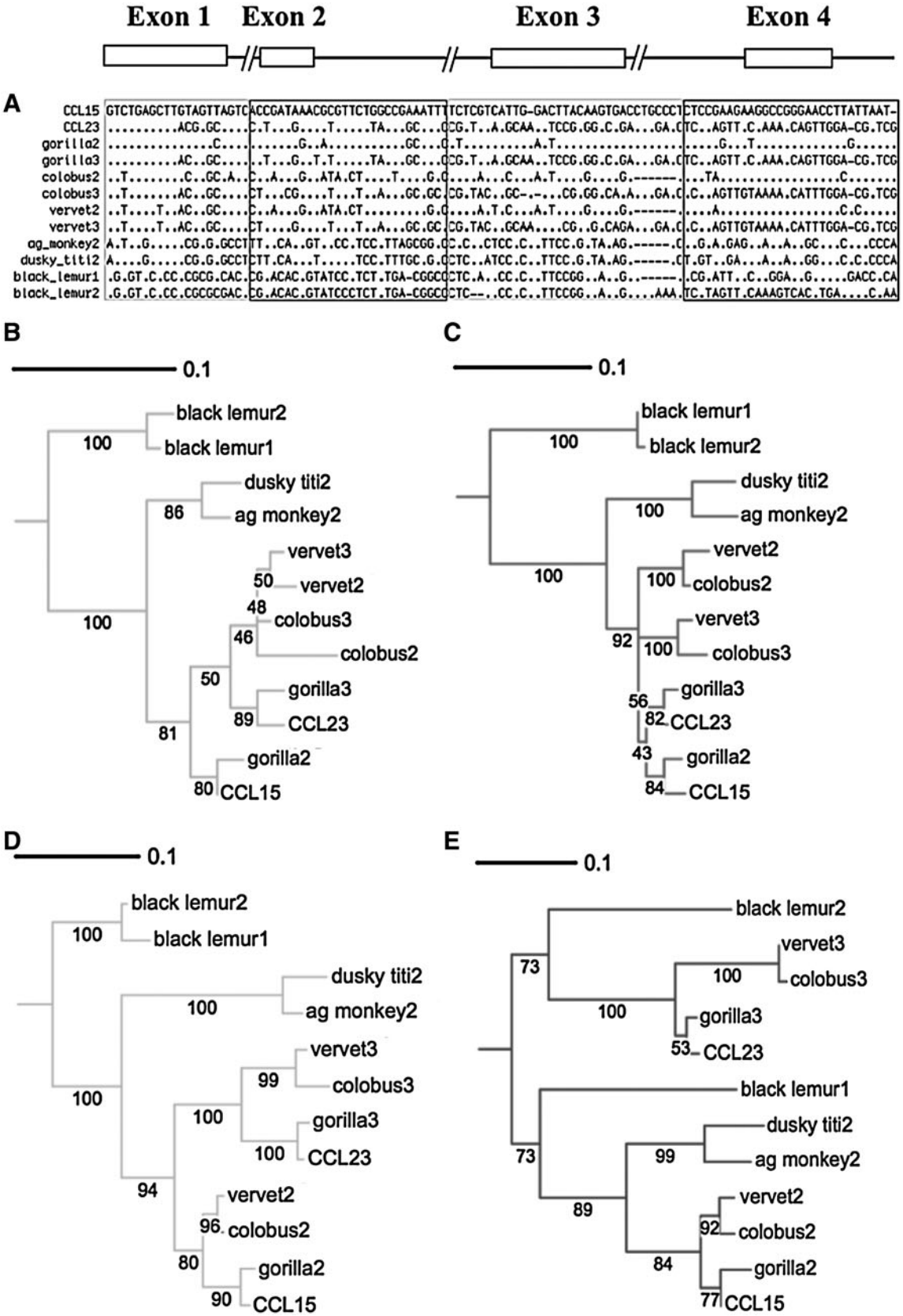

FIG. 10. Phylogenetic trees for four regions within the CCL15 and CCL23 genes in the primate CCL gene cluster. **(A)** Aligned sequences from the cluster, where only informative sites are shown. The alignment is divided into four regions based on our detected conversion events. **(B–E)** Maximum-likelihood phylogenetic trees (1000 bootstraps) are constructed for the regions containing exons 1–4, respectively.

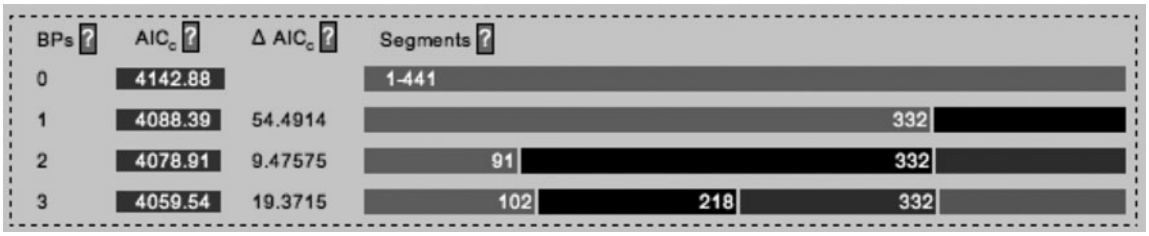

FIG. 11. Results from the GARD program (Pond et al., 2006). Evidence for three breakpoints is found, dividing the whole alignment into four regions near the boundaries of exons.

4. DISCUSSION

For much of the half-century since multi-gene families were discovered, it has been known that copies of the repeated genes within a species are often more similar than would be expected from their interspecies divergence. The processes generating this sequence homogeneity in repeated DNA are mechanisms of concerted evolution. Gene conversion is one of these processes, and while its impact on disease genes is appreciated (Chen et al., 2007), the extent of its impact on the evolution of the mouse genome has not been fully investigated in previous studies. Our work documents about one hundred and fifty thousand conversions (7.5%) between duplicated DNA segments in mouse. Similarly large fractions of conversion events among duplicated segments have been reported in whole-genome studies of yeast (Drouin, 2002), *Drosophila melanogaster* (Osada and Innan, 2008) and rodents (Ezawa et al., 2006), though the total number of observed gene conversions is much higher in our study. The genome-wide identification of DNA segments undergoing concerted evolution via gene conversions will make the application of comparative genomics to functional annotation considerably more accurate. This resource will allow the conversion

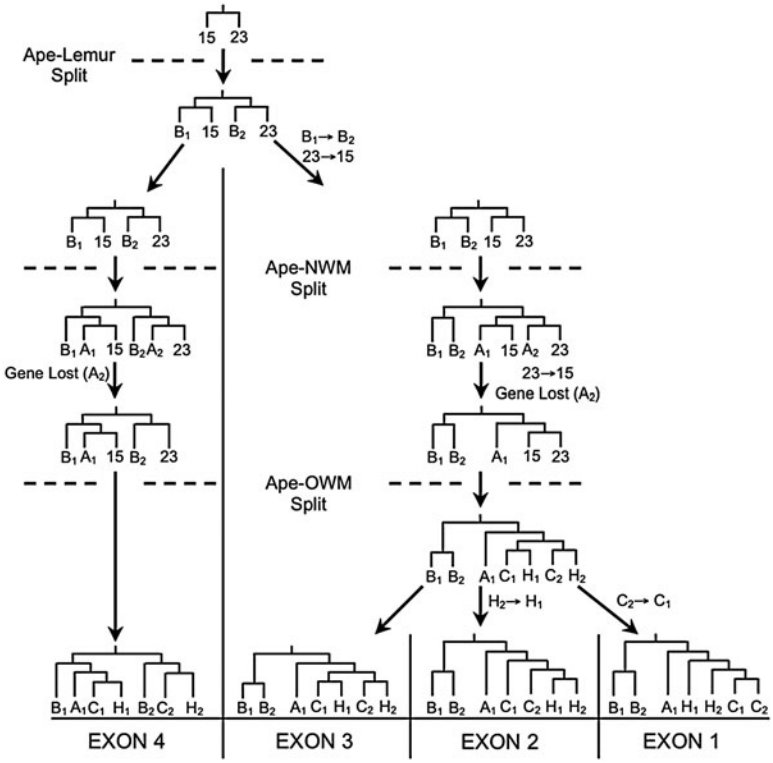

FIG. 12. Inferred evolutionary history of the primate CCL15 and CCL23 genes. Four different species, i.e., black lemur (B), howler monkey (A), colobus monkey (C), and human (H), are shown in these trees.

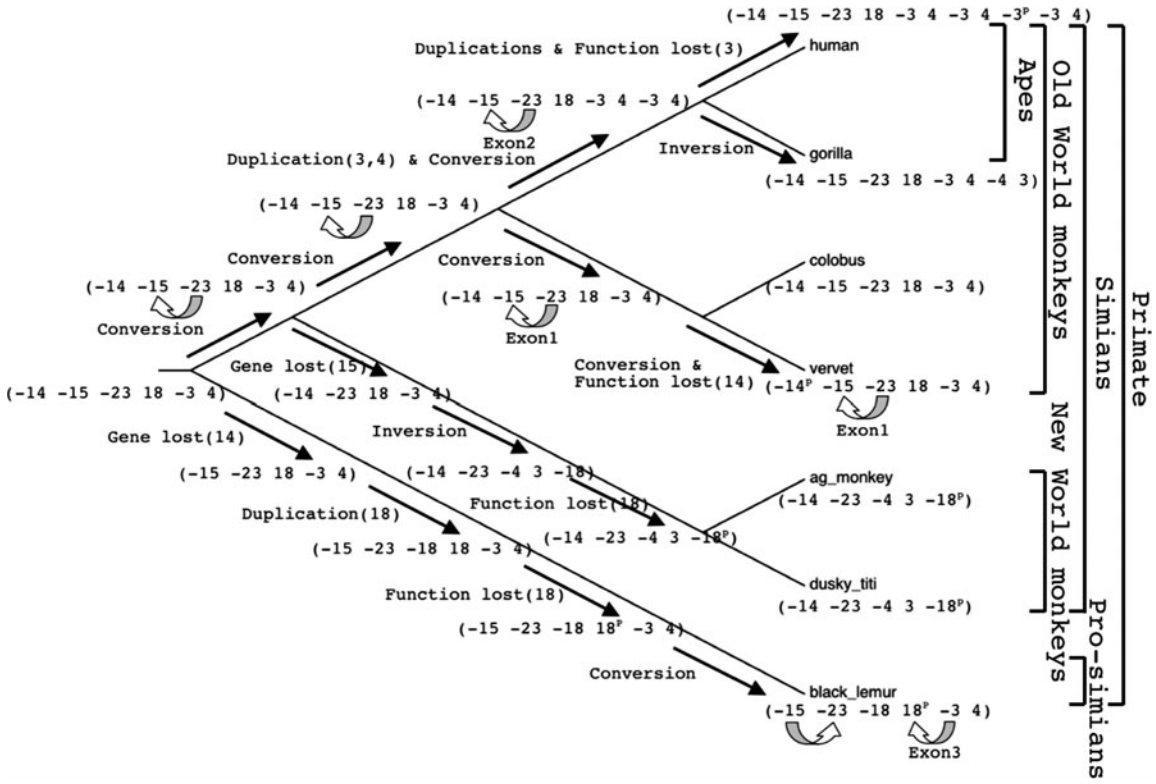

FIG. 13. Inferred evolutionary history of the CCL gene cluster. Here, superscript p indicates a pseudo-gene and "-" indicates a gene on the reverse strand.

process to be factored into functional inference based on sequence similarity to other species; for example, it could flag potential false positives for inferred positive or negative selection.

We also examined the association of gene conversion with various genomic features. In particular, we find that the length of paralogous segments has a strong positive effect on conversion events for both inter- and intra-chromosomal paralog pairs, as expected for a process requiring homologous pairing. Also, two findings indicate that closer proximity between the homologous pairs increases the likelihood of a conversion event: the conversion frequency is higher for intra-chromosomal pairs than for inter-chromosomal ones, and it is also higher for paralog pairs that are closer together on a chromosome. The closer proximity may be expected to increase the frequency of homologous pairing in recombination. Furthermore, the effects of coding sequences are very interesting. They have a positive correlation with conversion events, which could result from higher similarity. Also there is a more frequent conversion direction from a functional gene to a very similar pseudogene than vice versa, which could be a consequence of selection. The correlation with sequence similarity and the effects of GC content provide additional information about the occurrence of gene conversion events.

ACKNOWLEDGMENTS

This work was supported by grant HG02238 from the National Human Genome Research Institute to W.M., grant DK065806 from the National Institute of Diabetes, Digestive and Kidney Diseases to R.C.H., and funds provided by the Intramural Research Program of the National Human Genome Research Institute. We thank Maciek Boni for explaining his strategies for reducing the computer memory required by his method.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Agresti, A. 2002. *Categorical Data Analysis*. Wiley-Interscience, New York.

Benjamini, Y., and Hochberg, A. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* 85, 289–300.

Birney, E., Clamp, M., and Durbin, R. 2004. GeneWise and GenomeWise. *Genome Res.* 15, 988–995.

Boni, M., Posada, D., and Feldman, M. 2007. An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics* 176, 1035–1047.

Chen, J., Cooper, D., Chuzhanova, N., et al. 2007. Gene conversion: mechanisms, evolution and human disease. *Nat. Rev. Genet.* 8, 762–775.

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

Ezawa, K., Oota, S., and Saitou, N. 2006. Genome-wide search of gene conversions in duplicated genes of mouse and rat. *Mol. Biol. Evol.* 23, 927–940.

Feller, W. 1957. *An Introduction to Probability Theory and Its Application*. John Wiley & Sons, New York.

Felsenstein, J. 1989. PHYLIP—Phylogeny Inference Package (version 3.2). *Cladistics* 5, 164–166.

Holm, S. 1979. A simple sequential rejective multiple test procedure. *Scand. J. Statist.* 6, 65–70.

Hurles, M. 2004. Gene duplication: the genomic trade in spare parts. *PLoS Biol.* 2, 900–904.

Kent, W., Baertsch, R.A.H., Miller, W., et al. 2003. Evolutions cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. USA* 100, 11484–11489.

Kent, W., et al. 2002. The human genome browser at UCSC. *Genome Res.* 12, 996–1006.

Modi, W., et al. 2006. Genetic variation in the CCL18-CCL3-CCL4 chemokine gene cluster influences HIV Type 1 transmission and AIDS disease progression. *Am. J. Hum. Genet.* 79, 120–128.

Osada, N., and Innan, H. 2008. Duplication and gene conversion in the *Drosophila melanogaster* genome. *PLoS Genet.* 4 (12):e1000305.

Pond, S., Posada, D., Gravenor, M., et al. 2006. GARD: a genetic algorithm for recombination detection. *Bioinformatics* 22, 3096–3098.

Schwartz, S., Kent, W., Smit, A., et al. 2003. Human-mouse alignments with BLASTZ. *Genome Res.* 13, 103–107.

Semple, C., and Wolfe, K. 1999. Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. *J. Mol. Evol.* 48, 555–564.

Smit, A. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* 9, 657–663.

Xu, S., Clark, T., Zheng, H., et al. 2008. Gene conversion in the rice genome. *BMC Genom.* 9, 93–100.

Zhang, Z., Raghavachari, B., Hardison, R., et al. 1994. Chaining multiple-alignment blocks. *J. Comput. Biol.* 1, 217–226.

Address correspondence to:
Dr. Chih-Hao Hsu
Center for Comparative Genomics and Bioinformatics
Pennsylvania State University
University Park, PA 16802

E-mail: cxh503@psu.edu