Ectopic gene conversions in the human genome

David Benovoy¹, Guy Drouin*  
Département de biologie et Centre de recherche avancée en génomique environnementale, Université d'Ottawa, Ottawa, Ontario, Canada K1N 6N5

Article history:
Received 25 June 2008  
Accepted 18 September 2008  
Available online 19 October 2008

Keywords:
Ectopic  
Gene conversion  
Recombination  
Human genome  
Homo sapiens

A B S T R A C T

We used the GENCONV method to characterize the gene conversions that occurred amongst the 1434 protein coding human gene families with three or more genes. Conversions occur at a frequency of 0.88% (483 conversion events/55,050 gene pairs compared) and have an average length of 371±752 bp (±standard deviation). Both the size and the frequency of conversions are positively correlated with the similarity of the sequences involved in these conversions. The frequency of conversions and the local recombination rate are also positively correlated. Intrachromosomal conversions are almost 5 times more frequent than interchromosomal conversions and the frequency of intrachromosomal conversions increases as the distance between genes decreases. However, the higher frequency of conversions between nearby genes with the same transcriptional orientation is due to the fact that most functional duplicated genes are found next to one another and in the same transcriptional orientation. The average length of a conversion spanning only an intron region is significantly smaller than conversions spanning both exons and introns or only exons. This suggests that the smaller degree of sequence similarity of introns limits the size of conversions between duplicated human genes. The significant excess of conversions at the 3′-end of human genes suggests that incomplete cDNA molecules are often involved in conversions with chromosomal gene copies.

© 2008 Elsevier Inc. All rights reserved.

Introduction

The sequencing of the human genome allows us to apply a holistic approach to the study of the evolutionary processes that continually mold our genome. One such process is gene conversion, a process of non-reciprocal recombination that can either maintain genetic identity or promote genetic diversity [1]. Gene conversions are initiated by chromosomal double stranded breaks that are subsequently repaired by the strand invasion of a homologous sequence [2–5].

Conversions between alleles at the same locus are referred to as allelic gene conversions whereas conversion between dispersed sequences, found either on the same or on different chromosomes, are called ectopic gene conversions [6]. Gene conversions were first observed over 80 years ago, and then extensively studied, in yeast [7,8]. They have since been observed in a broad range of organisms from bacteria to humans [1,6,9–15]. In humans, gene conversions between multigene family members have been described in a wide variety of protein coding genes. For example, gene conversions have been shown to occur between human genes coding for β-globins, ubiquitins, and opsins [11,16–18]. The large palindromic sequences found in the human Y chromosome have also been shown to be subject to frequent gene conversion events [19].

Here, we use the GENECONV method [20] to identify ectopic gene conversions between duplicated human protein coding genes. We then analyze this data to explore the factors affecting ectopic gene conversions in the human genome and compare these characteristics with those of other organisms. In particular, we study the frequency of conversions, the average size of conversions, the sequence similarity of the regions flanking conversions, the proximity of converted genes, the transcriptional orientation of converted genes, the effect of introns sequences and the occurrence of conversions between genomic gene copies and their cDNAs.

Results

Frequency, size and flanking sequence similarity

We identified 483 ectopic gene conversion events in the data set produced by BLASTCLUST (which contained 1434 gene families with three or more members). The complete list of these 483 conversions as well as their chromosomal location, position of the conversions, length of the conversions, etc. is provided in Supplementary Table 1. The gene conversion frequency is 0.88% (483 conversion events/55,050 gene pairs compared) and there is a significant positive correlation (\( \rho^2 = 0.34, p = 5.32 \times 10^{-15} \); Spearman rank correlation test) between the number of gene conversions per gene family and the number of genes in each gene family.

The frequency of conversion events is not uniformly distributed throughout the human chromosomes. In fact, gene conversions occur...
more frequently in some regions than in others (Table 1). Furthermore, the frequency of conversions is positively correlated ($\rho^2=0.29$, $p=2.65 \times 10^{-12}$; Spearman rank correlation test) with variation in the local rate of recombination previously observed in the human genome (ref. 21; correlation calculated using windows of 5 Mb).

In protein coding genes, the average size of conversions is $371 \pm 752$ bp ($\pm$ standard deviation). The smallest gene conversion is $10$ bp long and the largest is $6011$ bp long. Fig. 1 describes the local rate of recombination previously observed in the human genome (ref. 21; correlation calculated using windows of 5 Mb).

In protein coding genes, the average size of conversions is $371 \pm 752$ bp ($\pm$ standard deviation). The smallest gene conversion is $10$ bp long and the largest is $6011$ bp long. Fig. 1 describes the relationship between the size of the gene conversion and the maximum sequence similarity computed between both genes involved in the gene conversion event in 100 nucleotides upstream or downstream from this event. Small conversion events (<1000 bp) occur between 23.8% and 100% DNA sequence similarity, whereas large conversion events (>1000 bp) only occur between 89% and 100% DNA sequence similarity. Furthermore, there is a highly significant positive correlation between the maximum sequence similarity of the flanking regions and the length of the gene conversion ($\rho^2=0.51$, $p=2.81 \times 10^{-27}$; Spearman rank correlation test). A weaker correlation is observed when comparing the minimum sequence similarity of the flanking regions with the length of the gene conversion ($\rho^2=0.31$, $p=1.18 \times 10^{-12}$; Spearman rank correlation test). Therefore, maximum sequence similarity was utilized in Fig. 1 because it is likely more representative of the molecular mechanism involved in gene conversion events. This is further illustrated in Fig. 2 where the maximum sequence similarity presents a strong positive correlation ($\rho^2=0.83$, $p=0.002$; Spearman rank correlation test) and the minimum sequence similarity present a non-significant correlation ($\rho^2=0.44$, $p=0.07$; Spearman rank correlation test) with the number of converted regions.

**Proximity and orientation**

Intrachromosomal conversions are almost 5 times more frequent than interchromosomal conversions (401 intrachromosomal versus 82 interchromosomal). The frequency of intrachromosomal gene conversion events also increases as the distance between genes implicated in the conversion event decreases. A Pearson correlation analysis performed on the ranked distances and ranked number of gene conversions shows a strong negative correlation ($\rho=-0.9$, $\rho^2=0.81$, $p=0.037$; Fig. 3). In addition, the median distance between intrachromosomal converted genes is $7.80 \times 10^4 \pm 2.88 \times 10^2$ (± standard deviation) bp. Given that the average gene density in the human genome is about 1 gene per $1 \times 10^5$ bp [22], this implies that the majority of converted genes are neighbors to one another. In fact, the

![Fig. 1. Relationship between the length of each converted region and the maximum similarity of its flanking regions. Maximum flanking similarity is the highest similarity value observed in either the 100 nucleotides upstream or downstream of each gene conversion event.](image)

![Fig. 2. Distribution of the number of gene conversions plotted against the maximum and minimum flanking similarity of the converted regions. Maximum flanking similarity is the highest similarity observed in either the 100 nucleotides upstream or downstream of each gene conversion event whereas minimum sequence similarity is the lowest of these two values.](image)

![Fig. 3. Distance between genes implicated in intrachromosomal gene conversion events, adjacent gene family members and all genes from the human genome. Note that in order to fit all data points on the same Y-axis, we divided the number of the “All genes” category by 10.](image)

### Table 1

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Bins with significant excess</th>
<th>$X^2$ value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>87.11</td>
<td>$6.17 \times 10^{-13}$</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>28.80</td>
<td>$7.00 \times 10^{-04}$</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>133.60</td>
<td>$2.86 \times 10^{-24}$</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>102.72</td>
<td>$4.42 \times 10^{-18}$</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>23.35</td>
<td>$5.46 \times 10^{-03}$</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>41.07</td>
<td>$4.86 \times 10^{-05}$</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>112.30</td>
<td>$5.01 \times 10^{-20}$</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>34.53</td>
<td>$7.21 \times 10^{-05}$</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>118.35</td>
<td>$2.90 \times 10^{-21}$</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>36.54</td>
<td>$3.17 \times 10^{-05}$</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>65.99</td>
<td>$9.26 \times 10^{-11}$</td>
</tr>
<tr>
<td>21</td>
<td>10</td>
<td>41.34</td>
<td>$4.33 \times 10^{-05}$</td>
</tr>
<tr>
<td>X</td>
<td>4</td>
<td>24.01</td>
<td>$4.28 \times 10^{-03}$</td>
</tr>
</tbody>
</table>

Notes. Only bins showing a significant excess ($p \leq 0.05$) of converted regions when compared to the expected value for a uniform distribution are listed. These bins are defined by dividing the length of a particular chromosome into 10 regions of equal lengths.
majority of converted genes are less than 1×10^4 bp (221 pairs of converted genes) and between 1×10^4 bp and 1×10^5 bp (90 pairs of converted genes) from each other (Fig. 3).

The 401 pairs of genes involved in the gene conversion events found on the same chromosome (intrachromosomal gene conversions) present a bias as to their orientation. Whereas the genes from families with three members or more are found in almost equal frequency (~50%) on the Watson and Crick strands [2657/5394 [49.2%] and 2737/5394 [50.7%], respectively), the genes involved in intrachromosomal gene conversions are found significantly more frequently on the same strand (277 pairs, 64.8%) than on opposite strands (124 pairs, 35.2%; \(\chi^2 = 21.16, p = 4.21 \times 10^{-4}\)).

Multigene family members found closely linked on the same chromosome are also found more frequently on the same strand than on opposite strands. Of the 179, 497 and 147 multigene family members found 0–1000, 1000–10,000 and 10,000–100,000 bp from one another, respectively, 163, 407 and 88 are one the same strand whereas 16, 90 and 59 are on opposite strands. There is therefore about 10, 4.5 and 1.5 times more genes on the same strand for genes found 0–1000, 1000–10,000 and 10,000–100,000 bp from one another, respectively. All these excesses are significant (\(p = 0, 0.02\) and 0.02, respectively). In contrast, of the 49, 51 and 8 multigene family members found 100,000 bp–1 Mb, 1–10 Mb and 10–100 Mb from one another, respectively, 25, 22 and 3 are one the same strand whereas 24, 29 and 5 are on opposite strands. Genes more than 100,000 bp from one another are therefore not significantly more often found on the same strand or on opposite strands (\(p = 0.89, 0.33\) and 0.48, respectively).

Effect of intron sequences

Fig. 4 shows that the length of converted intron regions, that is, the number of nucleotides of the conversion which originated in an exon that extends into an intron, is smaller than 100 bp before 75% maximum flanking sequence similarity but can be close to 900 bp long when the maximum flanking similarity is greater than 75%. The length of these converted intron regions and the maximum flanking similarity are significantly correlated (\(r^2 = 0.56, p = 6.96 \times 10^{-10}\); Spearman rank correlation test). The shortest and the longest converted intron regions are 2 and 904 bp long, respectively. The average converted intron region is 178±185 bp long (±standard deviation). This size is significantly smaller than the average length of a conversion spanning both an exon and an intron region (344±354 bp; \(W = 3526.5, p = 6.4 \times 10^{-8}\), Wilcoxon rank sum test).

Distribution of converted regions

The converted regions are not distributed randomly along the length of the converted genes (Fig. 5). Of the 966 converted regions, 199 were found in the last 10% of the genes. This is the only significant excess (\(\chi^2 = 108.5, p = 2.9 \times 10^{-19}\)) of converted regions detected in this distribution. All other bins show no significant deviation (chi-square tests results not shown) from the expected uniform distribution (96.6 converted regions per bin). Fig. 5 also shows the distribution of converted regions that cover both exon and intron sequences. For these converted regions, no excess in any region of the converted genes were detected (chi-square tests results not shown).

Discussion

The ectopic gene conversions found in the human genome share several characteristics with those found in other genomes. One such characteristic is the correlation between gene conversion lengths and flanking region similarity. Not only are long converted regions (>1000 bp) limited to regions where the flanking similarity is elevated (>89%; Fig. 1) but conversions are more frequent between more similar sequences (Fig. 2). This is similar to what has been observed in bacteria, yeast, Drosophila, C. elegans, mouse and human. In these species, the frequencies of allelic and ectopic gene conversion events have also been shown to depend on the degree of sequence similarity between the donor and recipient sequences [15,23–29]. For example, in Escherichia coli, a 2% mismatch between the donor and the recipient sequence can decrease the frequency of recombination four-fold and 10% mismatch can decrease recombination by over 40-fold [30,31]. These events are also dependent on the length of the sequences. In E. coli, there is an exponential increase in the frequency of recombination when the length of the sequences increases from 20 to 74 base pairs [30]. These results suggest that a high level of similarity in the flanking regions is necessary to stabilize and allow the migration of Holliday junctions. In contrast, low sequence similarity stops the migration of Holliday junctions and results in shorter conversions.

Interestingly, the 89% sequence similarity necessary to observe long conversions (Fig. 1) is similar to the 88% sequence similarity required for gene conversions to have an effect on the GC3-content of
dispersed duplicated yeast and Arabidopsis genes [32]. This suggests that a minimum of about 88% sequence similarity is necessary for gene conversions to have an important impact on the evolution of eukaryotic genomes.

Since gene conversions are the results of double-stranded breaks, their frequency is expected to be affected by the rate at which double-stranded breaks occur in different chromosomal regions. The study of Jensen–Seaman and colleagues [21] found that the rate of local recombination on a chromosome is not uniform but varies both within and between chromosomes. Using these local rates of recombination, we found that there is a significant correlation ($\rho^2 = 0.29$) between the frequency of gene conversion events and the rate of recombination. This implies that genes in regions where double-stranded breaks occur more often (recombination hot spots) are more likely to undergo gene conversions than genes that are in regions where double-stranded breaks occur less often (cold spots). Indeed, several human chromosomal regions have an excess of gene conversions (Table 1).

Another factor influencing ectopic gene conversion events is the distance between the pairs of converted genes. In the human genome, the frequency of gene conversion is inversely proportional to the distance between the pairs of genes involved in conversion events ($p = 0.9$). This observation is consistent with previous studies in yeast, C. elegans and mouse which also found a negative correlation between the frequency of gene conversions and the distance between gene pairs [15,23,27,33]. However, the higher frequency of conversions between nearby genes is due to the fact that most functional duplicated genes are found next to one another. In fact, if one normalizes the number of conversions in a given interval by dividing it by the number of adjacent gene family members in that interval (i.e., the numbers corresponding to the back and grey columns of Fig. 3, respectively), the relationship between distance and the (normalized) frequency of conversion is no longer significant ($p = 0.083$, Spearman rank correlation test). Thus, the higher frequency of conversions between nearby genes is simply the result of the fact that gene family members are most often found close to one another on the same chromosome. The fact that the chromosomal distribution of gene family members is different from that of unrelated genes is also reflected by the observation that the majority of gene family members are separated by less than 10,000 bp whereas most unrelated genes are separated by more than 10,000 bp (Fig. 3). The correlation between physical proximity and conversion frequency is therefore likely due to the close proximity of duplicated genes. This close proximity likely results from the fact that most functional duplicated genes originate from tandem duplications due to unequal crossing-overs between homologous chromosomes.

The observation that genes sharing the same transcriptional orientation convert almost twice as frequently as genes in opposite transcriptional orientation (64.8% versus 35.2%, respectively) is also most easily explained by the fact that duplication by unequal crossingovers will most often produce genes having the same transcriptional orientation. In fact, 80% (658 out of 823) of the multigene family members found less than 100,000 bp from one another on the same chromosome are on the same strand. The suggestion that this transcriptional orientation bias only reflects the arrangement of duplicated genes found on the same chromosome is also supported by the fact that this bias is not observed in the genes that are more than 100,000 bp from one another. Indeed, ~50% of these genes share the same transcriptional orientation whereas the other ~50% shares a different transcriptional orientation. These results are also consistent with those of Ezawa and colleagues [15] where it was shown that pairs of mouse genes with the opposite transcriptional orientation were as susceptible to gene conversions as pairs of mouse genes having the same transcriptional orientation.

In contrast with bacterial and most yeast (Saccharomyces cerevisiae) genes, which are essentially devoid of introns, human protein coding genes often contain introns. Our results show that these introns sequences have significant effects on conversions. In fact, the average length of converted intron sequence (178±185 bp) is significantly smaller than the average conversion lengths observed in exons and introns (344±354 bp, $p = 6.4 \times 10^{-8}$) or in exon sequences only (379±840 bp, $p = 6.1 \times 10^{-10}$). Our results therefore suggest that the lower similarity of intron sequences stops the elongation of converted regions. Furthermore, the relationship between the length of converted intron sequences and the similarity of the flanking region (Fig. 4) is similar to that of all converted regions (Fig. 1). Both of them show that long conversions only occur between very similar sequences.

Consistent with previous studies, our results suggest that cDNA molecules are often used as templates in the gene conversion process [23,34–37]. The distribution of converted regions within converted genes is uniform throughout the whole length of theses genes with the exception of the last 10% where there is a significant excess of converted regions (Fig. 5). This suggests that incomplete cDNA molecules are often used as a template during the conversion process. This hypothesis is supported by the distribution of converted regions in introns. The fact that there are no significant excesses of conversions in any gene regions when considering only conversions which extend into introns is consistent with this hypothesis because cDNA molecules do not contain intron sequences (Fig. 5). Incomplete cDNA molecules result from the poor processivity of reverse-transcriptases which initiate reverse-transcribe from the 3′-end and often fail to reverse-transcribe the full mRNA molecule [38]. Since such cDNA molecule lack intron sequences, they can only convert the 3′ coding regions thus creating the bias observed here. In contrast, conversions of intron sequences require that the conversions occur between two chromosomal sequences. In this case no region-specific bias is observed because all gene regions are equally likely to be involved in gene conversions.

Comparing the characteristics of human gene conversions with that of other organisms shows that some characteristics are common to distantly related organisms. For instance, a similar significant positive correlation between the number of gene conversions and the number of members in a gene family is observed in humans ($r^2 = 0.34$, $p = 5.32 \times 10^{-15}$), S. cerevisiae ($r^2 = 0.17$; ref. 23) and in four different E. coli genomes (K-12; $r^2 = 0.0028$, CFT073; $r^2 = 0.13$, EDL933; $r^2 = 0.15$, and Sakai; $r^2 = 0.11$; ref. 39). However, there are also several differences between the conversions observed in these organisms. First, human genes have a lower frequency of ectopic gene conversions. For humans we estimated it at 0.88%, a frequency about 9 times lower than those observed in S. cerevisiae (7.8%; ref. 23) and E. coli K-12 (7.6%; ref. 39). This difference could be due to the fact that human genes have long, and often very dissimilar, intron sequences whereas both yeast and E. coli genes are essentially devoid of intron sequences. The presence of (often very long and divergent) intron sequences in human genes likely decrease the frequency of gene conversions between duplicated genes because they decrease the similarity between duplicated genes and that high similarity is required for frequent conversions (Fig. 2). Second, the average size of a human gene conversion of 371±752 (±standard deviation) nucleotides is significantly longer than those found in S. cerevisiae (173±220 nucleotides; $W = 19719.5$, $p = 0.007$, Wilcoxon rank sum test; ref. 23) but significantly smaller than those found in E. coli K-12 (483±890 nucleotides; $W = 5556$, $p = 0.013$, Wilcoxon rank sum test; ref. 39). Third, conversions occur more frequently at the 3′-end of both human and yeast genes whereas they occur at the same frequency over the length of bacterial genes [28,39]. This suggests that the mechanisms employed to repair damaged DNA, and thus create a gene conversion, are different in different taxa.

Finally, we should mention that the 13% frequency of ectopic gene conversions reported by Ezawa and colleagues [15] for mouse genes (i.e., almost 15 times the 0.88% frequency we observed for human genes) is likely an overestimate due to the particular way in which they selected the mouse genes they analyzed. Since the test they used
to detect gene conversions is based on the analysis of quartets of sequences, their data set was only composed of sequences which were duplicated after the divergence of rodents and primates but before the speciation of mouse and rat. Since such recently duplicated sequences are more similar to one another than an unbiased sample of duplicated sequences, one expects to observe a higher frequency of gene conversions between them because, as discussed above, this frequency is positively correlated with sequence similarity. In contrast, our data set included all gene families containing at least three genes with sequences having at least 60% sequence similarity over at least 50% of their lengths.

The conclusions above assume that our results are not overly affected by false positives. This is a legitimate concern because the GENECONV method we used to detect gene conversions has been shown to give false positive results when the sequences being compared are more than 20% divergent [40]. To address this, we reanalyzed all our results by considering only the conversions that occurred between sequences having at least 80% sequence similarity. The only differences we observed in this second data set are that the frequency of conversions increased to 0.95% (324/34,042) and that the average size of conversions increased to 493±890 bp. Both these increases were expected given that our results (above) showed that conversions are more frequent and longer between more similar sequences. All other conclusions derived from this second data set were identical to those discussed above. For example, 1) there is a significant positive correlation between the number of gene conversions and the number of genes in each family (ρ2 = 0.46, p = 1.8 × 10−39; Spearman rank correlation test), 2) the frequency of conversion is positively correlated with the local rate of recombination (ρ2 = 0.36, p = 0.0017), 3) the chromosomal locations of the excesses of gene conversions were the same as those listed in Table 1, 4) the smallest gene conversion is 10 bp long and the largest is 6011 bp long, 5) the frequency of intrachromosomal conversions is negatively correlated with the distance between the converted genes (ρ = −1, ρ2 = 1, p = 0.0017), 6) the 284 pairs of genes involved in intrachromosomal gene conversions are found significantly more frequently on the same strand (187 pairs, 65.8%) than on opposite strands (97 pairs, 34.2%; χ2 = 14.26, p = 0.0002), and 7) the only significant excess of conversions occurs in the last 10% of the genes lengths (χ2 = 54, p = 1.8 × 10−9). Therefore, as in our recent study of bacterial genomes [28], considering all conversions or only conversions between genes having at least 80% sequence similarity between them does not affect our conclusions. Our conclusions are therefore robust in so far as they do not depend on the degree of similarity between the gene sequences analyzed.

The conclusions above also assume that the GENECONV method we used to detect gene conversions did indeed detect gene conversion events and not simply reflect the distribution of sequences differences between related sequences. At least one of our results can only be explained by the fact that the GENECONV method did detect bona fide conversion events, namely, the fact that the only significant excess of conversions occurs in the last 10% of the genes lengths. Clearly, this observation would not be expected if the GENECONV method had simply reflected the random distribution of sequences differences between related sequences. The conversion bias towards the 3′-end of coding regions is characteristic of gene conversions events which have been demonstrated experimentally in yeast (see above). Although no single gene conversion detection method can guarantee to detect all gene conversion events, our results therefore represent a valid characterization of the gene conversions present in the human genome [40–42].

In summary, this genome wide survey identified factors affecting gene conversions between human genes. These factors show both similarities and differences with the gene conversions observed in E. coli and yeast. In particular, the lower frequency of gene conversion in the human genome might be due to the lower recombination frequencies and to the presence of introns. It will be interesting to study gene conversions in other primate species in order to study their origin, fate and consequences.

Materials and methods

Sequences and multigene families

GenBank files (NCBI build 34.3) containing all known protein coding gene sequences (nucleotide and protein) and the 24 chromosomes sequences for Homo sapiens where downloaded from the NCBI ftp site (ftp://ftp.ncbi.nlm.nih.gov).

The BLASTCLUST program was downloaded from the NCBI ftp site and used to identify multigene families from the 27,350 known protein coding gene sequences. These gene families consisted in sequences having at least 60% sequence similarity over at least 50% of their lengths. The protein sequence similarity cut-off of 60% was chosen because a previous study found no extra gene conversions below this criterion [39]. Redundant gene copies due to alternative gene splicing were then removed and ClustalW [43] was used (with the default parameters) to align the protein sequences. A Perl script was used to align the corresponding DNA coding sequences to these protein alignments.

Gene conversion and statistical analyses

As in previous studies from our group, the GENECONV 1.7 program (http://www.math.wustl.edu/~sawyer/geneconv/) was used to identify gene conversions within the human genomes [23,39]. Using the same method as with other genomes (S. cerevisiae and E. coli, respectively) allows us to compare the general characteristics of their gene conversions with the ones found in this study of the human genome. Significant (p ≤ 0.05) local inner fragments with some mutation (g-scale = 2) were used. Duplicate gene conversions were removed from the data set using phylogenetic analyses of each gene family as previously described [23].

Previous studies have assessed the power (type II errors) and the rate of false positives (type I errors) of 14 different recombination detection methods using simulated and empirical data sets [40,41]. The GENCONV method performed well with both types of data when using gene families with at least 3 members, had a type I error rate of about 5% and were relatively free of type II errors (false negatives) when the sequence divergence was ≥ 25%. This method is more reliable when using gene families containing three or more members because it accounts for polymorphic sites and thus allow the differentiation between true gene conversions and mutation cold or hot spots [20,23].

Converted regions bordering introns, or spanning multiple exons, were further investigated. If any gene conversions where found to be near an exon–intron boundary (≤ 2 nucleotides away), the intron(s) sequences of this pair of genes where aligned and then inserted into the corresponding exon–intron boundary (ies) of the aligned coding sequences of this pair of genes. This new alignment was then used as input for the GENECONV program. The g-scale value was set to 2 and the “Include_monosites” option was used because it permits the analyses of only two sequences. Since analyzing only two sequences removes the control for conserved sites (such as shared repetitive elements), any significant gene conversions detected inside an intron sequence was further investigated. These converted intron sequences were blasted against the ALU and the non-redundant database at NCBI and, if any significant hits for repeated elements were detected for a particular intron sequence, that gene conversion event was removed from the analysis. This allowed us to remove spurious intron sequence similarity due to the presence of similar repeated sequence in introns.

All statistical analyses were performed using S-plus v6.2 (In-sightful Corporation, Seattle, WA) and Excel (Microsoft Corporation, Redmond, WA).
Distance, orientation and location of duplicated and converted genes

The chromosomal location of each converted gene and conversion event was calculated relative to the first nucleotide in the chromosome sequence file. The orientation of the pair of genes involved in the conversion event was established by aligning, using FASTA34 [44], these genes and determining if the reverse complement of the gene in question was utilized (reverse strand) or not (forward strand) to align it with its corresponding chromosome sequence. In addition, to verify if the orientation of converted genes is biased, we compared them to the orientation of all paralogous genes found in the human genome. The distance between intrachromosomal converted genes was computed by using the difference between the center positions of both converted gene regions.

Acknowledgments

We would like to thank both anonymous reviewers for their useful comments. This work was supported by a Discovery Grant from the National Science and Engineering Research Council of Canada to G. D.

Appendix A. Supplementary data


References