Two Patterns of Genome Organization in Mammals: the Chromosomal Distribution of Duplicate Genes in Human and Mouse

Robert Friedman and Austin L. Hughes
Department of Biological Sciences, University of South Carolina, Columbia

Gene duplication occurs repeatedly in the evolution of genomes, and the rearrangement of genomic segments has also occurred repeatedly over the evolution of eukaryotes. We studied the interaction of these two factors in mammalian evolution by comparing the chromosomal distribution of multigene families in human and mouse. In both species, gene families tended to be confined to a single chromosome to a greater extent than expected by chance. The average number of families shared between chromosomes was nearly 60% higher in mouse than in human, and human chromosomes rarely shared large numbers of gene families with more than one or two other chromosomes, whereas mouse chromosomes frequently did so. A higher proportion of duplicate gene pairs on the same chromosome originated from recent duplications in human than in mouse, whereas a higher proportion of duplicate gene pairs on separate chromosomes arose from ancient duplications in human than in mouse. These observations are most easily explained by the hypotheses that (1) most gene duplications arise in tandem and are subsequently separated by segmental rearrangement events, and (2) that the process of segmental rearrangement has occurred at a higher rate in the lineage of mouse than in that of human.

Introduction

In a recent review, Eichler and Sankoff (2003) cite localized duplication of genomic segments and rearrangement of chromosomal segments as two major factors in eukaryotic genome evolution. Despite the availability of draft genome sequences from a number of eukaryotes, it remains unclear how these factors have interacted over evolutionary history to yield different genomic structures in different groups of organisms. Gene duplication is important in adaptive evolution because it makes possible the evolution of new proteins having distinct functions (Li 1982; Hughes 1994, 1999). However, it is so far unknown to what extent segmental rearrangements merely randomly reshuffle the products of gene duplication and to what extent the physical rearrangement of duplicated genes is important for their functional differentiation.

Analysis of complete genome sequences from eukaryotes suggests that gene duplication, giving rise to multigene families, occurs continually over evolution (Lynch and Conery 2000). Whereas most duplicate genes are eventually lost, certain duplicate genes are retained, and a duplicate gene that assumes a new function selectively advantageous to the organism is more likely to be retained than one of neutral effect (Hughes 1994; Lynch et al. 2001). The genomes of mammals include genes that have been duplicated throughout evolutionary history. Multigene families now present in mammalian genomes have been shown by phylogenetic analysis to include duplicates that originated before the origin of vertebrates, duplicates that originated early in vertebrate history, duplicates that originated in the mammalian lineage before the radiation of the eutherian (placental) orders, and duplicates that originated after the radiation of the eutherian orders (Friedman and Hughes 2001, 2003; Gu, Wang, and Gu 2002). Analyses of the human genome have shown that more recently duplicated genes are more likely to be physically linked than are more anciently duplicated genes, apparently reflecting the predominance of tandem duplication as a mechanism of gene duplication (Friedman and Hughes 2003).

Comparison of genetic maps of different species of mammals has suggested that, over the evolution of mammals, genomes have evolved by rearrangement of genomic segments, including syntenic groups of genes (O’Brien et al. 1999; Band et al. 2000). The recent completion of a draft sequence of the human and mouse genomes provided further support for this view (Mouse Genome Sequencing Consortium 2002). Although it has so far not been possible to reconstruct with certainty the ancestral arrangement of syntenic groups in mammals (Mouse Genome Sequencing Consortium 2002), the availability of genomic sequence has improved reconstruction of the breakpoints involved in the rearrangements that took place since the last common ancestor of human and mouse (Pevzner and Tesler 2003). The rearrangement of genomic segments provides a mechanism whereby genes originally duplicated in tandem can be spread to different chromosomes over the course of evolution (Friedman and Hughes 2003).

In the present paper, we address the question of how the rearrangement among chromosomes over mammalian evolutionary history has affected the distribution of duplicate genes. Examining a set of gene families present in both human and mouse genomes, we compare their chromosomal distributions in the two species. Our results show that these two mammalian species have very different overall patterns of distribution of gene families among chromosomes, which suggests, in turn, that the chromosomal redistribution of multigene family members may be a factor in the evolutionary differentiation of different lineages of eukaryotes.

Methods

Sequence Data

The genomic data for human (version 16.33) and mouse (version 16.30) were obtained from Ensembl (http://www.ensembl.org). The version numbers refer to...
the database software (Ensembl version 16) and the assembly of the genomic sequences (NCBI versions 33 and 30). Genes predicted by Ensembl have been curated and verified by similarity with homologs discovered experimentally (Clamp et al. 2003). The numbers of annotated protein-coding genes were 32,035 for human and 32,911 for mouse. After removal of genes that were shorter than 100 bases and longer than 300,000 bases, the total gene sets were 29,606 in human and 32,296 in mouse. Further curation to remove proteins encoded by alternative splicing variants at the same locus resulted in a count of 20,387 genes in human and 23,222 genes in mouse.

Protein families were identified by homology and a single-linkage method employed by the BlastClust software available in the Blast software package (Altschul et al. 1997). Sequence homology was established by identifying matches using a conservative E-value of $10^{-6}$ with a minimum of 30% sequence identity across at least 50% of length of two sequences. Preliminary analyses with this data set and others showed that this set of criteria is conservative for establishing gene family membership (Hughes and Friedman 2004; unpublished data). The single-linkage method assembles larger families by linking shared genes among families, thus ensuring that a given gene will be assigned to only one family. Parsing and processing of sequence and gene family data were performed by software written in Perl.

Evolutionary Analyses

Given family membership and chromosomal location information, we compared the distribution of families across chromosomes in the two species. In these analyses, we excluded the Y chromosome, which has only a small number of genes. For a set of families having at least two members in each of the two species, we compared the distribution of families on the autosomes and the X chromosome with the random expectation by a randomization test. This test was conducted by creating simulated genomes in which the chromosomal locations of the genes analyzed were randomly reassigned. In each simulated genome, both the set of genes and the set of gene locations on chromosomes were the same as in the real genome, but all genes were randomly assigned to chromosomal locations. The distribution of genes in the real genome was compared with that in 1,000 simulated genomes.

Homologous sequences were aligned at the amino acid level using the ClustalW program (Thompson, Higgins, and Gibson 1994), and this alignment was imposed on the DNA sequences. The number of synonymous nucleotide substitutions per synonymous site ($d_s$) and the number of nonsynonymous nucleotide substitutions per nonsynonymous site ($d_N$) were estimated by a maximum-likelihood method (Yang and Nielsen 2000) using the software package PAML (Yang 1997).

Results

Distribution of Gene Families

We identified 7,104 gene families present in both human and mouse genomes. The mean number of genes per family in mouse (2.52) was slightly but significantly greater than the mean number of genes per family (2.27) in human (table 1). When we excluded families present in a single copy in both of the genomes, the mean number of genes per family in mouse (4.91) remained significantly higher than that in human (4.25) (table 1).

Considering only the subset of families present in at least two copies in a given genome (2,260 families in human and 2,322 families in mouse), we compared the distribution of families across the autosomes and the X chromosome with the random expectation. In both species, the number of families confined to a single chromosome was significantly greater than expected by chance (table 2). In human, the number of families found on just two chromosomes was slightly fewer than expected by chance, and the difference was significant at the 5% level (table 2). By contrast, in mouse, the number of families found on just two chromosomes was not significantly different from the random expectation (table 2). In both species, the numbers of families found on three or more chromosomes was significantly fewer than expected by chance (table 2).

Thus, in comparison to randomly rearranged genomes, both human and mouse showed a tendency toward more families confined to a single chromosome and fewer multiple-chromosome families. This tendency appears to be indicative of the predominance of tandem duplication as the major mechanism of gene duplication in both species (Friedman and Hughes 2003).

When we examined the pairwise numbers of families shared between chromosomes within each of the two species, we found that the median number shared was significantly higher in mouse than in human (fig. 1). In addition, the mouse genome showed far more cases where two chromosomes shared 100 or more gene families than did the human genome (fig. 2). In the human genome, there were 12 chromosomes (out of 22 autosomes plus X) that shared at least 100 families with at least one other chromosome (fig. 2A). When sharing of 100 or more families among these 12 chromosomes was represented as a graph, the graph was fully connected; but the total number of connections was surprisingly low. The minimal number of connections for a fully connected 12-node graph is 11, and the graph of human chromosomes sharing 100 or more families had only 15 connections. Only chromosome 1 (eight connections), chromosome 12 (four connections), and chromosome 2 (three connections) had more than one or two connections in this graph (fig. 2A).

In the mouse, 14 chromosomes (of 19 autosomes plus X) shared 100 or more families with at least one other chromosome (fig. 2B). When sharing of 100 or more families among chromosomes in the mouse genome was represented as a graph, this graph also was fully connected (fig. 2B). However, the observed number of connections was 46, which is over three times the minimum number (13) for a fully connected 14-node graph. Of the 14 chromosomes, only two (chromosomes 6 and 15) shared 100 or more families with as few as two other chromosomes. Chromosome 2 shared at least 100 families with 13 other chromosomes, and both chromosomes 7 and 11 shared at least 100 families with 11 other chromosomes (fig. 2B).
Nucleotide Substitution

To obtain evidence regarding the relative time of duplication of duplicate genes in human and mouse, we estimated the number of synonymous nucleotide substitutions per synonymous site \((d_S)\) and the number of nonsynonymous nucleotide substitutions per nonsynonymous site \((d_N)\) in comparisons between the members of two member families within each species. A total of 2,349 comparisons between such gene pairs were made (1,156 in human and 1,193 in mouse). For both species, mean \(d_S\) for gene pairs on different chromosomes was higher than that for gene pairs on the same chromosome (fig. 3). This difference was highly significant \((P < 0.001)\) in a factorial analysis of variance using species and location on the same or different chromosomes as main effects. Mean \(d_S\) for comparisons between gene pairs on the same chromosome was higher in mouse than in human, but mean \(d_S\) for comparisons between gene pairs on different chromosomes was higher in human than in mouse (fig. 3). This effect was supported by a highly significant interaction \((P = 0.001)\) between species and location on the same or different chromosomes in the analysis of variance. By contrast, the analysis showed no significant difference between species. When a similar analysis was applied to \(d_N\), no significant effects were observed (not shown).

An examination of the distribution of \(d_S\) values shed further light on these results. For both species, the distribution of \(d_S\) values for both within-chromosome and between-chromosome comparisons was trimodal. These three modes possibly correspond to three main "peaks" of gene duplication in vertebrate history. The first mode \((d_S < 1)\) evidently corresponds to duplications that have occurred since the rodent-primate divergence (dated at 110 MYA [Kumar and Hedges 1998]). The second mode \((1 < d_S < 3)\) may correspond to duplications in the tetrapod lineage before the mammalian radiation. The third mode \((3 < d_S < 5)\) may correspond to duplications early in vertebrate history, before the origin of the tetrapods.

The relative magnitude of these three modes differed strikingly, depending on the species and on whether the genes compared were on the same or different chromosomes. For human gene pairs on the same chromosome, the first mode was by far the most prominent, with the third mode being barely detectable (fig. 4A). On the other hand, for mouse genes on the same chromosome, the first and second modes were about equally prominent, and the third mode was more clearly visible than in the human case (fig. 4B). The relative prominence of the second and third modes in the case of mouse gene pairs on the same chromosome explains why mean \(d_S\) for within-chromosome comparisons was higher in mouse than in human (fig. 3).

### Table 1
Mean Numbers of Genes Per Family (± SE) in Mapped Portions of Human and Mouse Genomes

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>Mouse</th>
<th>(P^a)</th>
<th>(P^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete data set ((N = 7,104))</td>
<td>2.27 ± 0.19</td>
<td>2.52 ± 0.23</td>
<td>0.004</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Families with at least two members in at least one of the species ((N = 2,772))</td>
<td>4.52 ± 0.48</td>
<td>4.91 ± 0.59</td>
<td>0.004</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

\( ^a \) Paired t-test.
\( ^b \) Wilcoxon signed rank test.

### Table 2
Numbers of Human and Mouse Gene Families with Two or More Members Mapped to Different Numbers of Chromosomes and Comparison with 1,000 Random Genomes

<table>
<thead>
<tr>
<th>Number of Chromosomes</th>
<th>Human Observed random genomes</th>
<th>1</th>
<th>2</th>
<th>≥3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>49.8</td>
<td>1174.9</td>
<td>1022.3</td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>7.6</td>
<td>10.8</td>
<td>8.2</td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>35–84</td>
<td>1141–1207</td>
<td>995–1047</td>
<td></td>
</tr>
<tr>
<td>(P)</td>
<td>&lt;0.001</td>
<td>&lt;0.05</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>Mouse Observed random genomes</td>
<td>200</td>
<td>1209</td>
<td>913</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>67.9</td>
<td>1202.0</td>
<td>1052.1</td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>7.8</td>
<td>10.7</td>
<td>7.6</td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>45–94</td>
<td>1167–1237</td>
<td>1028–1077</td>
<td></td>
</tr>
<tr>
<td>(P)</td>
<td>&lt;0.001</td>
<td>NS</td>
<td>&lt;0.001</td>
<td></td>
</tr>
</tbody>
</table>
The pattern for comparisons of gene pairs on different chromosomes was nearly the opposite of that for within-chromosome comparisons. In human, the third mode was substantially more prominent than the other in between-chromosome comparisons (fig. 4C). In mouse, on the other hand, although all three modes were clearly present, the first mode had a greater prominence than in human, and the third mode had reduced prominence in comparison with human (fig. 4D). These differences between the two species explain why, for between-chromosome comparisons, mean $d_S$ was higher in human than in mouse.

**Discussion**

Gene duplication has been hypothesized to be a repeated occurrence in the evolution of genomes (Lynch and Conery 2000; Friedman and Hughes 2003), and there is evidence that rearrangement of genomic segments has also occurred repeatedly over the evolution of eukaryotes (O’Brien et al. 1999; Eichler and Sankoff 2003). We studied the interaction of these two factors in mammalian evolution by comparing the chromosomal distribution of multigene families in human and mouse. The results revealed both similarities and differences between the two species.

In both species, gene families tended to be confined to a single chromosome to a greater extent than expected by chance (table 2). This pattern of gene distribution probably reflects, at least in part, the fact that many gene families arise by tandem duplication and remain linked in tandem arrays. This same biological fact also apparently explains why, in both genomes, fewer families are found on multiple (>3) chromosomes (table 2).

On the other hand, the distribution of families across chromosomes was strikingly different between the two species. The average number of families shared between chromosomes was nearly 60% higher in mouse than in human (fig. 1). Note that this difference is too large to be a consequence simply of the slightly (11%) larger average family size in mouse than in human (table 1). Human chromosomes rarely shared large numbers of gene families with more than one or two other chromosomes, whereas mouse chromosomes frequently did so (fig. 2).

Comparison of the patterns of sharing of large numbers (>100) of gene families in the two species highlighted the unique nature of human chromosome 1. This chromosome shared 100 or more gene families with...
eight other chromosomes, whereas no other human chromosome shared 100 or more gene families with more than three other chromosomes (fig. 2A). There is evidence that human chromosome 1 represents an ancestral chromosome in eutherian (placental) mammals that has been independently rearranged in different eutherian lineages, although retained in primates (Murphy et al. 2003). If so, the role of this chromosome as a kind of superchromosome sharing large numbers of gene families with numerous others presumably also reflects a genomic pattern ancestral to eutherian mammals.

The maximum-likelihood (ML) method (Yang and Nielsen 2000) provides estimates of the number of synonymous substitutions per site ($d_S$) even when it would be impossible to estimate that quantity by a simpler method. This method yields estimates even when multiple changes are hypothesized to have occurred at each site; thus, the higher values of $d_S$ estimated by this method should probably be treated with some caution. However, some authors (e.g., Blanc, Hokamp and Wolfe 2003) have argued that ML estimates of $d_S$ are meaningful even when $d_S$ is substantially greater than one substitution per site.

In the present data set, it seems likely that the relative magnitude of $d_S$ estimates contains biological meaning, although the sharp separation between the three observed modes (fig. 4) may partly be an artifact of the estimation process. Nonetheless, the trimodal distribution of $d_S$ in comparisons between duplicated genes in human and mouse (fig. 4) is of interest because the three modes appear to correspond to three periods of gene duplication in the mammalian lineage that have been identified by phylogenetic methods; namely, duplications after the mammalian radiation, duplications within the tetrapod lineage before the mammalian radiation, and duplications in early vertebrates before the origin of tetrapods (Friedman and Hughes 2001, 2003; Gu, Wang, and Gu 2002).

Fig. 4.—Frequency distributions of $d_S$ in human and mouse between the members of two-member gene families located on the same and different chromosomes.
Our analyses revealed differences between human and mouse regarding the chromosomal distribution of duplicate gene pairs that appear to have originated during these three periods of gene duplication. In human, pairs on the same chromosome were found to correspond mainly to recent duplications, whereas those on separate chromosomes corresponded to more ancient duplications (fig. 4). This pattern is consistent with phylogenetic analyses showing that, in the human genome, within-chromosome duplicates disproportionately result from recent duplications, whereas between-chromosome duplications disproportionately result from ancient duplications (Friedman and Hughes 2003). This, in turn, is consistent with a model whereby most gene duplication occurs by tandem duplication, and over long periods of evolutionary time, tandem duplicates are eventually separated by chromosomal rearrangements (Friedman and Hughes 2003).

In comparison with human, the mouse genome showed a much higher proportion of ancient duplicates on the same chromosome and of recent duplicates on different chromosomes. One possible explanation for this difference is that the human chromosomal arrangement is closer to that of the eutherian ancestor than is that of the mouse and that segmental rearrangements have occurred to a greater extent in the mouse lineage than in the human lineage. Segment rearrangements may have dispersed recent duplicates throughout the genome to a greater extent in human than in mouse. By the same token, they may have brought back onto the same chromosome certain ancient duplicates that had been separated in the eutherian ancestor. A greater rate of interchromosomal segmental exchange in the rodent than in the primate lineage may further explain the fact that, in the mouse, a much higher proportion of chromosomes share large numbers of gene families (figs. 1 and 2), because this latter situation might also be a result of a more extensive mixing of ancestral mammalian chromosomal segments in the rodent than in the primate lineage.

Acknowledgment

This research was supported by grant GM066710 to A.L.H. from the National Institutes of Health.

Literature Cited


Two Patterns of Genome Organization in Mammals 1013

William R. Jeffery, Associate Editor

Accepted December 10, 2003