Sequences of Primate Insulin Genes Support the Hypothesis of a Slower Rate of Molecular Evolution in Humans and Apes than in Monkeys

Susumu Seino,* Graeme I. Bell,* ‡ and Wen-Hsiung Li‡

*Howard Hughes Medical Institute, Department of Medicine, and ‡Department of Biochemistry and Molecular Biology, The University of Chicago; and ‡Center for Demographic and Population Genetics, University of Texas at Houston

The chimpanzee and African green monkey insulin genes have been cloned and sequenced. These two sequences together with the previously reported sequences for the human and owl monkey insulin genes provide additional support for the hominoid-rate-slowdown hypothesis, i.e., a slower rate of nucleotide substitution in humans and apes than in monkeys. When these sequences and other primate sequences available for the relative-rate test were considered together, the substitution rate in the Old World monkey lineage was shown to be significantly higher than the rates in the human and chimpanzee lineages. This was true regardless of whether the η-globin pseudogene was included in the analysis. Therefore, in contrast to the claim by Easteal, the hominoid-rate-slowdown is not unique to the η-globin pseudogene but appears to be a rather general phenomenon. On average, the substitution rate at silent sites is about 1.5 times higher in the Old World monkey lineage than in the human and chimpanzee lineages.

Introduction

There has been a longstanding controversy over the hominoid-rate-slowdown hypothesis, which postulates that the rate of molecular evolution has become slower in hominoids after their separation from the Old World (OW) monkeys. This hypothesis, first proposed by Goodman (1961) and Goodman et al. (1971), is based on rates estimated from immunological distance and protein sequence data. Sarich and Wilson (1967) and Wilson et al. (1977) contended that the slowdown was an artifact, owing to the use of an erroneous paleontological estimate of the ape-human divergence time. They conducted relative-rate tests using both immunological distance data and protein sequence data and concluded that there was no evidence for a hominoid slowdown. A similar conclusion was drawn from DNA hybridization studies (Kohne et al. 1972; Sibley and Ahlquist 1984); however, more recent DNA hybridization studies have produced conflicting conclusions (Sibley and Ahlquist 1987; Caccone and Powell 1989). On the other hand, comparative analyses of DNA sequence data by Britten (1986), Li and Tanimura (1987), and Li et al. (1987) provided strong support for the hominoid-slowdown hypothesis, and so the hypothesis has been accepted by many molecular evolutionists. However, recently Easteal (1991) has argued that the slowdown occurred only in the η-globin pseudogene, because, when this pse-

1. Key words: nucleotide substitution, molecular clocks, hominoid slowdown, primate evolution.

Address for correspondence and reprints: Wen-Hsiung Li, Center for Demographic and Population Genetics, University of Texas, P.O. Box 20334, Houston, Texas 77225.

0737-4038/92/0902-0001$02.00
dogene was removed from comparison, the rate of nucleotide substitution in the OW monkey lineage was no longer significantly higher than that in the human lineage. To resolve this controversy, more DNA sequence data are needed.

In our studies of the molecular evolution of the insulin gene, we have determined the nucleotide sequence of the insulin gene from the owl monkey (Aotus trivirgatus), a New World (NW) monkey (Seino et al. 1987). Here, we report the sequences of the chimpanzee and African green monkey (an OW monkey) insulin genes. These three primate sequences together with the human insulin gene sequence (Bell et al. 1980; Ullrich et al. 1980) are used to test the hominoid-slowdown hypothesis. We show below that the new data support the slowdown hypothesis. Thus, in contrast to the claim by Easteal (1991), the phenomenon is not unique to the η-globin pseudogene.

Material and Methods

DNA

Chimpanzee (Pan troglodytes) DNA was isolated from fibroblast cultures (GMO3448A) obtained from the NIGMS Human Genetic Mutant Cell Repository, Coriell Institute for Medical Research, (Camden, N.J.). African green monkey (Ceropithecus aethiops) DNA was isolated from CV-1 cells (CCL70; American Type Culture Collection, Rockville, Md.).

Southern Blotting

DNA was digested with EcoRI, electrophoresed, blotted, and hybridized with the 32P-labeled human insulin gene probe phins96 (Bell et al. 1984), according to a method described elsewhere (Bell et al. 1981).

Insulin Gene Isolation and Sequencing

Genomic DNA was digested with EcoRI and fragments of 10–15 kb were isolated by electrophoresis and cloned in λCh4A (Maniatis et al. 1982, pp. 270–294). Phage containing the chimpanzee and African green monkey insulin genes were identified by cross-hybridization with the human insulin gene (Seino et al. 1987).

The EcoRI fragments containing the chimpanzee and African green monkey insulin genes were subcloned into pUC9. Subfragments of these EcoRI fragments were cloned into M13mp19 and sequenced by the dideoxy chain-termination method (Sanger et al. 1980). More than 85% of the sequence was determined on both strands. The nucleotide sequence was compared with that of the human insulin gene (Bell et al. 1980, 1982; Ullrich et al. 1980). Sites at which the sequences differed were checked to confirm that the differences were not due to sequencing or clerical errors.

Data Analysis

The sequences of the chimpanzee and African green monkey insulin genes are from the present study. The other insulin gene sequences were obtained from Bell et al. (1980, 1982), Ullrich et al. (1980), Wetekam et al. (1982), and Seino et al. (1987) (GenBank accession numbers J02989 and J00268). For the sources for the other sequences studied, see the legend to table 3.

For coding regions, we use the method of Li et al. (1985) to compute the number of substitutions per nonsynonymous site ($K_\alpha$) and the number of substitutions per synonymous site ($K_S$) between two sequences. For noncoding regions we use Kimura's (1980) two-parameter method to estimate the number of substitutions per site ($K$). For the relative-rate test (Sarich and Wilson 1967), we use the statistics developed by
Wu and Li (1985). Sequence alignment was simple and was done manually; the regions of tandem repeats in the 5’ flanking region were excluded from comparison because they differ greatly in length and sequence among species.

**Results**

**Sequences of Chimpanzee and African Green Monkey Insulin Genes**

The human insulin gene hybridizes to a single EcoRI fragment in many primate species whose size is about 12–13 kb (S. Seino and G. I. Bell, unpublished observations). EcoRI fragments containing the chimpanzee and African green monkey insulin genes were cloned, and the sequences of regions of 2,483 and 1,909 bp, respectively, determined (fig. 1). The boundaries of the exons were assigned by comparison with the human insulin gene sequence (Bell et al. 1980; Ullrich et al. 1980). The exon-intron organization of the insulin genes of these two primates is the same as those previously reported for the human and owl monkey insulin genes (Bell et al. 1980; Ullrich et al. 1980; Seino et al. 1987) and corresponds to what is termed the “ancestral” insulin gene structure (Steiner et al. 1985), i.e., the coding region of the gene being interrupted by two introns. This is different from the rat I and mouse I insulin genes, which have only a single intron located in the region of the gene encoding the 5’ untranslated region of the mRNA and which are believed to have been generated by a process of retrotransposition of a partially processed insulin transcript (Soares et al. 1985).

The identity between the sequences of human (chimpanzee) and African green monkey preproinsulin is 98% (97%); at the nucleotide level the corresponding identity between these protein-coding regions is 98% (95%). The amino acid sequence of human, chimpanzee, and African green monkey insulin is identical. Chimpanzee preproinsulin (fig. 1A) differs from the human protein at two sites in the signal peptide: amino acids -13 and -2 are Ala in the human protein. The sequence of African green monkey preproinsulin (fig. 1B) differs at amino acids -3 and -2 in the signal peptide and at residue 37 in the C-peptide portion of the precursor, which are Val, Pro, and Leu, respectively. In addition, there are several other notable differences between the human, chimpanzee, and African green monkey insulin gene sequences. The first is a deletion of 48 bp in the chimpanzee insulin gene at a site just after the TAG which specifies termination of translation (fig. 1). As a consequence, the predicted 3’ untranslated region of chimpanzee mRNA is only 28 nucleotides, exclusive of the poly(A) tract and may be the shortest of any of the insulin mRNAs described to date if the site at which it is polyadenylated is the same as that for human insulin mRNA. This deletion is unlikely to be a cloning artifact, since it was a feature of eight independently isolated clones. The deletion of this 48-nucleotide segment in the chimpanzee insulin gene implies that this region is not required for translation of the mRNA.

The second notable difference was in the 5’ flanking region of these genes. The chimpanzee insulin gene has a region of 206 bp that begins −365 bp from the start of transcription and that comprises 20 tandem repeats of sequences related to the 15-bp sequence ACAGGGGTCTCTGGGG (fig. 1). A region of similar tandem repeats is present in the human gene (Bell et al. 1982), the most common sequence of which, ACAGGGGTCTCTGGGG, does not occur in the chimpanzee gene. However, the next two most common repeats in the human are ACAGGGGTCTCTGGGG and ACAGGGGTCTCTGGGG, which constitute most of the repeats in the chimpanzee. This region is characterized by extensive polymorphism in humans (Bell et al. 1981) and thus may be polymorphic in chimpanzees as well. Similar regions of multiple GC-rich tandem repeats are not present in the African green monkey or owl monkey
B. African Green Monkey

Exon 1

Intron A

Exon 2

Exon 3

FIG. 1.—Sequences of chimpanzee and African green monkey insulin genes and flanking regions. The boundaries of the genes and the nucleotide which is polyadenylated in the mRNA were assigned by comparison with the human insulin gene. The predicted start of transcription has been designated as nucleotide 1. The deduced amino acid sequences of the chimpanzee and African green monkey preproinsulins are shown. The signal peptide is amino acids -24 to -1; the B chain, 1 to 30; C-peptide (including pairs of adjacent basic amino acids), 31 to 65; and A chain, 66 to 86. The sequences of the introns are shown in lowercase letters. The region of tandem repeats in the 5'-flanking region of the chimpanzee insulin gene is shown in boldface type. A similar region is not present in the African green monkey insulin gene, and the sequence which is in its place is shown in boldface type. The number of the nucleotide or amino acid is shown in italics at the beginning of each line. The arrow following the amber (TAG) codon of the chimpanzee insulin gene indicates the location of a 48-bp deletion relative to the human insulin gene. The polyadenylation signal, AATAAA, is underlined. The sequences have been submitted to the GenBank™/EMBL Data Bank under accession numbers X61089 and X61092.
insulin genes. Rather, the insulin gene of the African green monkey has in its place a sequence, ACAGGGGTCCCCAGGACAGGGGG TCTGGGG, which is homologous to two copies of the consensus repeats found in the human and chimpanzee genes. The owl monkey has the sequence GCAGGGGTCTGGGG in this region (Seino et al. 1987). Thus, these regions of tandem repeats may only be a feature of the insulin genes of humans and higher apes.

Sequence Divergence between Species

Table 1 shows the number of nucleotide substitutions per 100 sites in noncoding regions between species; the 5' and 3' untranslated regions are not included because they are short. In the following comparisons, the data compiled by Li et al. (1987) and Easteal (1991) will be used as references. First, we consider the 5' flanking region. The divergence between human and chimpanzee is 2.1%, which is slightly higher than the average divergence (~1.7%) in noncoding regions and at synonymous sites between human and chimpanzee sequences (see Li et al. 1987; Easteal 1991). The divergence (9.5%) between human (or chimpanzee) and the African green monkey is at the higher end of the variation observed among other genes. The divergence (10.7%) between human and the owl monkey is close to that (11.1%) observed for the η-globin pseudogene (between the same two species), and the divergence (14.3%) between the African green monkey and the owl monkey is also close to that (13.3%) observed in the η-globin pseudogene.

Next, we consider the two introns. The divergences (1.3% and 2.0%) in both introns between human and chimpanzee are within the variation observed among other genes. In intron B the divergence (8.6%) between human and the African green monkey is within the variation observed among the introns in other genes, whereas in intron A the divergence (13.7%) is higher than that observed in other genes. The divergences in introns A and B (15.3% and 16.1%, respectively) between human and the owl monkey are considerably higher than those observed in other genes and in the η-globin pseudogene. In the comparison between the African green monkey and the owl monkey the divergence is relatively low (11.7%) in intron A but relatively high (18.2%) in intron B, when compared with the divergence in the η-globin pseudogene. Finally, the 3' flanking region tends to show a higher degree of divergence.

Table 1
Mean ± Standard Error of Number of Nucleotide Substitutions per 100 Sites, between Species, in Noncoding Regions of Insulin Gene

<table>
<thead>
<tr>
<th>Species Pair</th>
<th>5' Flanking Region (395 bp)</th>
<th>Intron A (160 bp)</th>
<th>Intron B (699 bp)</th>
<th>3' Flanking Region (93 bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human-Chimpanzee</td>
<td>2.1 ± 0.7</td>
<td>1.3 ± 0.9</td>
<td>2.0 ± 0.5</td>
<td>2.8 ± 1.6</td>
</tr>
<tr>
<td>Human-African green monkey</td>
<td>9.5 ± 1.6</td>
<td>13.7 ± 3.2</td>
<td>8.6 ± 1.2</td>
<td>13.0 ± 4.2</td>
</tr>
<tr>
<td>Human-owl monkey</td>
<td>10.7 ± 1.8</td>
<td>15.3 ± 3.4</td>
<td>16.1 ± 1.7</td>
<td>21.6 ± 5.3</td>
</tr>
<tr>
<td>Chimp-African green monkey</td>
<td>9.5 ± 1.6</td>
<td>13.7 ± 3.2</td>
<td>8.2 ± 1.2</td>
<td>13.0 ± 4.2</td>
</tr>
<tr>
<td>Chimp-owl monkey</td>
<td>11.9 ± 1.9</td>
<td>15.3 ± 3.4</td>
<td>15.4 ± 1.6</td>
<td>23.0 ± 5.5</td>
</tr>
<tr>
<td>African green monkey-owl monkey</td>
<td>14.3 ± 2.1</td>
<td>11.7 ± 2.9</td>
<td>18.2 ± 1.8</td>
<td>23.6 ± 6.0</td>
</tr>
</tbody>
</table>
than do other regions, though the higher values are probably partly due to large standard errors.

Table 2 shows $K_S$ and $K_A$. The $K_S$ value of 5.1% between human and chimpanzee is much higher than the values observed in other genes (see Li et al. 1987; Easteal 1991). This is also the case for the $K_S$ values between human (or chimpanzee) and the three monkey species. The $K_S$ value [17.5% (or 19.1%)] between the African green monkey (*Macaca fascicularis*) and the owl monkey is within the variation observed among other genes. The $K_A$ values between human, chimpanzee, the African green monkey, and *M. fascicularis* are small, reflecting the fact that there are only a few nonsynonymous differences in both the signal peptide and the C peptide regions but no nonsynonymous differences in the regions coding for the A and B chains (see Seino et al. 1987). On the other hand, the $K_A$ values between the owl monkey and the other four species are relatively large, because several amino acid differences have been found between the owl monkey insulin polypeptide and the insulin polypeptides from the other four species (see Seino et al. 1987).

**Evidence for a Hominoid Slowdown**

The insulin gene sequences are useful for examining the hominoid-slowdown hypothesis. We use the relative-rate test. For example, consider the African green monkey as species 1, human as species 2, and the owl monkey as species 3, i.e., a reference. Let $K_{ij}$ denote the number of substitutions between species $i$ and $j$. Then, a positive sign for $K_{13} - K_{23}$ means that the African green monkey lineage evolved faster than the human lineage, and a negative sign means the opposite. It is seen from table 3 that in the introns and the flanking regions the African green monkey lineage has evolved faster than the human lineage. On the other hand, at synonymous sites the human lineage has evolved faster than the African green monkey lineage; this occurred because, as mentioned above, the human and chimpanzee sequences have evolved exceptionally fast at synonymous sites. [The $K$ values at synonymous sites differ slightly from those reported by Li et al. (1987) because the African green monkey sequence was used instead of that of *M. fascicularis*.] The chimpanzee insulin gene sequence shows a pattern similar to that in the human sequence, i.e., compared with the African green monkey sequence, it has evolved faster at synonymous sites but slower in the introns and flanking regions. Overall, the insulin gene sequences provide some evidence for a rate slowdown in hominoids (humans and apes).

**Table 2**

*Mean ± Standard Error of Number of Substitutions per 100 Synonymous Sites (above Diagonal) and per 100 Nonsynonymous Sites (below Diagonal)*

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>Chimpanzee</th>
<th>African Green Monkey</th>
<th>Cynomolgus Monkey</th>
<th>Owl Monkey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chimpanzee</td>
<td>5.1 ± 2.6</td>
<td>19.5 ± 5.6</td>
<td>17.7 ± 5.2</td>
<td>25.0 ± 6.6</td>
<td></td>
</tr>
<tr>
<td>African green monkey</td>
<td>0.8 ± 0.6</td>
<td>16.6 ± 5.2</td>
<td>14.9 ± 4.7</td>
<td>21.7 ± 6.1</td>
<td></td>
</tr>
<tr>
<td>Cynomolgus monkey</td>
<td>1.3 ± 0.7</td>
<td>1.7 ± 0.9</td>
<td>9.2 ± 3.6</td>
<td>17.5 ± 5.6</td>
<td></td>
</tr>
<tr>
<td>Owl monkey</td>
<td>6.6 ± 1.7</td>
<td>7.0 ± 1.8</td>
<td>7.0 ± 1.8</td>
<td>6.6 ± 1.7</td>
<td></td>
</tr>
</tbody>
</table>
Table 3
Differences in Number of Nucleotide Substitutions per 100 Sites, between Old World
Monkey (Species 1) Lineage and Human or Chimpanzee (Species 2) Lineage

<table>
<thead>
<tr>
<th>SEQUENCE (N*)</th>
<th>HUMAN</th>
<th>CHIMPANzee</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$K_{12}$</td>
<td>$K_{12} - K_{23}$</td>
</tr>
<tr>
<td>Synonymous:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insulin (28)</td>
<td>19.5</td>
<td>$-7.4 \pm 6.3$</td>
</tr>
<tr>
<td>$\alpha$-Globin (106)</td>
<td>6.1</td>
<td>1.1 ± 3.4</td>
</tr>
<tr>
<td>$\beta$-Globin (114)</td>
<td>23.0</td>
<td>$-1.0 \pm 8.3$</td>
</tr>
<tr>
<td>$\delta$-Globin (74)</td>
<td>8.9</td>
<td>2.8 ± 5.6</td>
</tr>
<tr>
<td>$\gamma$-Globin (105)</td>
<td>11.5</td>
<td>$-1.1 \pm 4.1$</td>
</tr>
<tr>
<td>Transforming growth factor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alcohol dehydrogenase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apolipoprotein Al (158)</td>
<td>7.9</td>
<td>$-5.3 \pm 4.8$</td>
</tr>
<tr>
<td>Apolipoprotein E (228)</td>
<td>10.6</td>
<td>5.1 ± 3.2</td>
</tr>
<tr>
<td>Erythropoietin (145)</td>
<td>11.2</td>
<td>5.1 ± 5.9</td>
</tr>
<tr>
<td>Pepsinogen (263)</td>
<td>10.5</td>
<td>0.0 ± 3.3</td>
</tr>
<tr>
<td>$\alpha$-Antitrypsin (140)</td>
<td>10.9</td>
<td>6.7 ± 6.8</td>
</tr>
<tr>
<td>Pro-opiomelanocortin (180)</td>
<td>15.6</td>
<td>$-6.2 \pm 4.2$</td>
</tr>
<tr>
<td>Introns:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insulin (859)</td>
<td>9.6</td>
<td>1.0 ± 1.3</td>
</tr>
<tr>
<td>$\alpha$-Globin (237)</td>
<td>4.5</td>
<td>$-4.5 \pm 2.1^*$</td>
</tr>
<tr>
<td>$\beta$-Globin (190)</td>
<td>8.2</td>
<td>3.2 ± 3.5</td>
</tr>
<tr>
<td>$\beta$-Globin (960/915)</td>
<td>7.4</td>
<td>2.1 ± 1.2</td>
</tr>
<tr>
<td>$\delta$-Globin (863)</td>
<td>9.8</td>
<td>$-0.4 \pm 1.1$</td>
</tr>
<tr>
<td>Flanking and untranslated regions:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insulin (348)</td>
<td>9.8</td>
<td>3.2 ± 1.5$^*$</td>
</tr>
<tr>
<td>$\alpha$-Globin (195)</td>
<td>5.0</td>
<td>2.4 ± 2.3</td>
</tr>
<tr>
<td>$\delta$-Globin (402)</td>
<td>10.2</td>
<td>3.4 ± 3.5</td>
</tr>
<tr>
<td>$\beta$-Globin (146/179)</td>
<td>3.5</td>
<td>1.7 ± 1.8</td>
</tr>
<tr>
<td>$\delta$-Globin (338)</td>
<td>6.3</td>
<td>1.0 ± 1.6</td>
</tr>
<tr>
<td>Transforming growth factor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pro-opiomelanocortin (230)</td>
<td>15.6</td>
<td>4.7 ± 3.9</td>
</tr>
<tr>
<td>Overall (7,409/2,655)</td>
<td>9.3</td>
<td>1.1 ± 0.4$^{b, **}$</td>
</tr>
</tbody>
</table>

NOTE.—In all comparisons the first species is an Old World monkey and the second species is human (first case) or chimpanzee (second case). The insulin sequences are from the present study (the African green monkey and chimpanzee) and GenBank; the owl monkey is used as the third (reference) species. For chimpanzee $\beta$-globin, the OW monkey species used is Macaca cynomolgus and the reference species is brown lemur; all sequence data are from GenBank. All other results are taken from Li et al. (1987) and Easteal (1991). $K_{ij}$ = no. of substitutions, per 100 sites, between species $i$ and $j$.

* Approximate number of sites compared; since a site in the coding region can be partially synonymous and partially nonsynonymous, $N$ is often not an integer, but for convenience we have used integers.

$^b$ Computed by using $1/(K_{13} - K_{23})$ as weight.

$^* P < 0.05$ (if the mean $\geq$ 2 times the standard error).

$^{**} P < 0.01$ (if the mean $\geq$ 2.7 times standard error).

In the above test, we have used only the insulin gene sequences. We now also include other sequences. Since Easteal (1991) claims that a hominoid slowdown is observed only in the $\eta$-globin pseudogene sequence, we shall exclude this sequence from analysis. One way to compute the mean and variance of $K_{13} - K_{23}$ when all available sequences are considered together is to combine the sequences in tandem.
and treat them as a single sequence. In the present case, however, this approach is not applicable, because the reference species (i.e., species 3) used is not the same for all sequences. We have therefore computed the "overall" mean and variance by using the reciprocal of the variance of $K_{13} - K_{23}$ as weight. This test is more conservative, since it tends to have a smaller chance of rejecting the null hypothesis of equal rate. Even so, it is clear from table 3 that, when all the sequences used are considered together, the substitution rate in the OW monkey lineage is significantly higher than the rates in the human and chimpanzee lineages (see bottom of table 3). Thus, even with the exclusion of the $\gamma$-globin pseudogene sequence, there is evidence for the hominoid-slowdown hypothesis. Note that the sequences used are the same as those used by Easteal (1991), except that we have added the flanking sequences and the introns of the insulin gene.

Discussion

The controversy over the molecular-clock hypothesis has been continuing unabated since the time of its proposal by Zuckerkandl and Pauling (1965). One major difficulty in resolving this controversy is that molecular changes are subject to strong stochastic effects, so that many data are needed to show a significant difference in rate. In Easteal's (1991) study, the molecular-clock hypothesis could not be rejected if the $\gamma$-globin pseudogene was excluded from comparison. Now, with the addition of new data, the molecular-clock hypothesis can be rejected. Therefore, the hominoid slowdown is not unique to the $\gamma$-globin pseudogene but seems to be a rather general phenomenon in the average sense. However, the fact that $K_{13} - K_{23}$ is negative in many cases in table 3 suggests that some sequences have indeed evolved faster in the hominoid lineages than in the OW lineage, and so the hominoid slowdown probably does not occur in all genes.

In the present study a slower rate in the human lineage than in the OW monkey lineage has been interpreted as a slowdown in the former lineage rather than as a speedup in the latter lineage, because the rate in the human lineage also appears to be slower than those in other mammalian lineages such as rodents and artiodactyls (see Britten 1986; Li et al. 1987). Also, this interpretation fits better the generation-time-effect hypothesis (Kohne et al. 1972), which postulates that the rate of nucleotide substitution is higher in short-living organisms than in long-living ones because the number of DNA replications per unit time (and so also the rate of mutation) in the germ line would be higher in the former than in the latter.

As nucleotide substitutions are subject to strong stochastic effects, the rate difference between two lineages is likely to vary from sequence to sequence. Therefore, many data may be required for obtaining a reliable estimate of the rate difference between two lineages. For the sequences used to compare the rates in the OW monkey lineage and the human lineage, the $K$ values for the two lineages are $(9.3\% + 1.1\%)/2 = 5.2\%$ and $9.3\% - 5.2\% = 4.1\%$, respectively; therefore, the OW monkey lineage has, on average, evolved 1.3 ($=5.2/4.1$) times faster than the human lineages. For the sequences used to compare the rates in the OW monkey lineage and the chimpanzee lineage, the $K$ values are 4.75% and 3.25%, respectively, and so the former lineage has evolved 1.5 times faster than the latter lineage. This ratio of 1.5 is actually the same as the ratio between the two lineages for the $\gamma$-globin pseudogene ($\sim 10,000$ bp), a ratio which can be computed from the $K$ values in Easteal's (1991) study. Note that this ratio refers to the ratio of the average rates along each lineage from the common ancestor of the two lineages to the present time. Since at the early stage of divergence
the two lineages would have very similar rates of nucleotide substitution, the rate ratio at the present time would be considerably higher than the long-term average. Therefore, the ratio at the present time may be of the same order as that between the generation times in OW monkeys and humans, which is probably between two and three.

Acknowledgments

This study was supported by the Howard Hughes Medical Institute and by NIH grants DK20595 (to G.I.B.) and GM30998 and GM45670 (to W.-H.L.).

LITERATURE CITED


WESLEY M. BROWN, reviewing editor

Received June 27, 1991; revision received September 17, 1991

Accepted September 17, 1991
Mitochondrial DNA–like Sequence in the Nuclear Genome of an Akodontine Rodent

Margaret F. Smith,* W. Kelley Thomas,† and James L. Patton*

*Museum of Vertebrate Zoology, and †Division of Biochemistry and Molecular Biology, University of California, Berkeley

Initial amplification and sequencing of a 366-bp fragment of the cytochrome b gene by a conserved primer pair (MVZ 03 and MVZ 04) revealed a nonfunctional copy of the gene with two deletions (one of which is 17 bp in length and the other of which is 3 bp in length) in Chroeomys jelskii, a South American akodontine rodent. By means of an alternative primer to MVZ 03—namely, MVZ 05—from the region of the tRNA for glutamic acid, a functional copy of cytochrome b was subsequently amplified. Both primer pairs amplify functional sequence when applied to purified mitochondrial DNA (mtDNA). Restriction-endonuclease digestion of purified mtDNA from C. jelskii did not reveal any additional sets of bands that would suggest heteroplasmy in the mitochondrial genome. When probed with both functional and nonfunctional gene fragments, MboI restriction digests revealed the same pattern, providing further evidence that the nonfunctional copy must be located in the nucleus. Observed differences in the mitochondrial and nuclear sequences from two populations are consistent with a faster rate of change in mtDNA than in nuclear DNA.

Introduction

Mitochondrial DNA (mtDNA) fragments have been reported in the nuclear genome of humans (Fukuda et al. 1985; Nomiyama et al. 1985; Kamimura et al. 1989), rats (Hadler et al. 1983), snow geese (Quinn and White 1987), sea urchins (Jacobs and Grimes 1986), locusts (Gellissen et al. 1983), maize (Kemble et al. 1983), fungi (Wright and Cummings 1983), and yeast (Farrelly and Butow 1983; Thorsness and Fox 1990). We report here on an mtDNA-like segment in the nuclear genome of one species of South American akodontine rodent, Chroeomys jelskii.

Material and Methods

DNA was extracted from frozen liver tissue from four individuals of Chroeomys jelskii by using the sodium dodecyl sulfate–proteinase K/phenol/RNase method (Maniatis et al. 1982, pp. 458–462). Purified mtDNA for two of the four individuals of C. jelskii was obtained from frozen liver and kidney tissue by using cesium chloride–ethidium bromide gradient centrifugation (Lansman et al. 1981). Liver extracts from three other species of akodontine rodents—including Akodon andinus, A. boliviensis,
Nuclear Copy of a Mitochondrial DNA Sequence 205

and *Bolomys amoenus* (for additional details about these other species, see Smith and Patton 1991)—were used in comparisons with *C. jelskii* in the present report.

DNA sequence was obtained for segments of the cytochrome *b* gene (*cyt* *b*) by using four primers referred to here by their number in a series from the Laboratory of the Museum of Vertebrate Zoology (MVZ) (University of California, Berkeley). The primer sequences are as follows: MVZ 03 (L14230), 5'-GCTTCATCCAACATCTCAGT-3'; MVZ 07 (L14230), 5'-AACCCCATCTAACATTCCTCWTGATG-3'; MVZ 05 (L14115), 5'-CGAAGCCTTGATATGAAAAACCATCGTTG-3'; and MVZ 04 (H14542), 5'-GCAGCCCCTCAGAAATATTTGGCCTC-3'. The letters in parentheses identify the light (i.e., L) or heavy (i.e., H) strand of mtDNA, and the numbers following the letters give the position of the 3’ base of the primer in the complete mtDNA sequence for the house mouse (Bibb et al. 1981).

Various combinations of primers were used (fig. 1). MVZ 03 or MVZ 07 in combination with MVZ 04 should amplify a 311-bp segment of the *cyt* *b* gene. With the primers added at each end, the total length of the amplified fragment should be 366 bp. MVZ 05 in combination with MVZ 04 amplifies a 426-bp segment (the initial bases are from the tRNA for glutamic acid, followed by sequence from the *cyt* *b* gene); the addition of the primers at each end gives a total length of 482 bp.

The general procedures used for amplifications and sequencing have been detailed by Smith and Patton (1991). Amplifications from liver and purified mtDNA extracts were successful under a variety of conditions. All of the liver extracts from *C. jelskii* eventually gave double-stranded products with primer pair MVZ 03–MVZ 04 at an annealing temperature of 60°C, but a reduced annealing temperature of 45°C resulted in products for more individuals in a given reaction. For the purified mitochondrial extracts an annealing temperature of 45°C was used most often, but good yields of double-stranded products were also produced at annealing temperatures of 50°C and 55°C.

Restriction digests of the purified mtDNA were made using restriction enzyme *MboI* according to the manufacturer's instructions. The fragments were subsequently separated in a 2% agarose gel and were transferred to nitrocellulose by following standard procedures (Maniatis et al. 1982, pp. 382–389). Radioactive probes for the nonfunctional (nuclear) and functional (mitochondrial) copies were generated by the polymerase chain reaction (PCR) by following the same amplification protocol, except that 10 μCi P32 dCTP was included. Hybridization of filters with labeled probes was in 5 X SSPE (180 mM NaCl, 10 mM NaH2PO4, pH 7.4), 0.5% SDS at 65°C for 16 h. Filters were subsequently washed in 0.1 X SSPE, 0.5% SDS for 1 h at 65°C.

----

![Fig. 1.](image-url)

Fig. 1.—Location of primers used to amplify various *cyt* *b* fragments in South American akodontine rodents. The numbering is from the complete mtDNA sequence for the house mouse (Bibb et al. 1981).
Sequences were entered into BIONET (IntelliGenetics 1988) for translation and alignment, with the alignments subsequently being adjusted by eye. Restriction-digest simulations were also done with BIONET.

The sequences for 401 bp of the mtDNA cyt b gene for 12 species of South American akodontine rodents, including *C. jelskii*, have been described by Smith and Patton (1991) and are available from GenBank (accession numbers M35691—M35716). MVZ catalog numbers are given below for the voucher specimens of the four individuals of *C. jelskii*.

*Chrooemys jelskii*: Peru: Depto. Junin; 22 km N La Oroya, 4040 m (MVZ 173083, 173084). Peru: Depto. Puno; 6.5 km SW Ollachea, 3350 m (MVZ 173073, 173074).

**Results and Discussion**

Detection and Characterization of a Nonfunctional cyt b Sequence

In all four individuals of *Chrooemys jelskii* the only double-stranded product produced using primer pair MVZ 03–MVZ 04 was smaller (fig. 2) than those in 11 other species in the *Akodon* group. The anomalous cyt b sequences obtained from *C. jelskii* by primer pair MVZ 03–MVZ 04 (fig. 3) have a 17-bp deletion, resulting in a frameshift, and an additional 3-bp deletion. There are three chain-termination codons within the anomalous sequence. Among the missing amino acids is one of the histidines considered to be essential in binding the heme group in a functional cyt b molecule (Howell and Gilbert 1988). All of the above evidence indicates that the cyt b sequence obtained from *C. jelskii* by primer pair MVZ 03–MVZ 04 is nonfunctional.

The primer combination MVZ 05–MVZ 04 did yield functional sequences for
FIG. 3.—Sequence for maximum of 305 bp of cyt b gene for four *Chroeomys jelskii* individuals, two each from two geographic localities. The sequence for this region from the house mouse, beginning at base 14235 (see Bibb et al. 1981), is given first. Next is the sequence obtained with primer pair MVZ 05—MVZ 04 (functional copy). The final lines in each block are the sequence obtained with primer pair MVZ 03—MVZ 04 (nonfunctional copy; X = deletion). For the nonfunctional sequences the bands on the sequencing gels were not strong enough to permit scoring either the first 21 bp from individual MVZ 173084 or the first 24 bp from MVZ 173073. Y = Sites in the functional sequence where bands of equal strength appear to be present in both the C and T lanes; we were unable to resolve the ambiguity. Only one individual from each locality was used for the comparisons with Mus in table 1.

all individuals of *C. jelskii* (fig. 3). There are no frameshifts or stop codons in these sequences.

Probable Reason for the Behavior of the Primers

Because the fragment amplified by the primer combination MVZ 05—MVZ 04 includes the area of the gene where the primer MVZ 03 anneals, we were able to determine the degree of mismatch between the MVZ 03 primer and its complement on the functional gene. The primer MVZ 03 is 27 bases long, including four bases at the 5' end of the primer that are part of an added restriction site and that are not counted as mismatches here. The MVZ 03 primer fails to match the functional sequence in *C. jelskii* at seven sites, compared with four mismatches each in *Bolomys amoenus* and in *A. aerosus* from one locality and compared with fewer mismatches in other species (see examples in fig. 4). Apparently, in *C. jelskii*, the primer MVZ 03 attaches preferentially to a nonfunctional copy of the gene. On the other hand, the primer MVZ 05 anneals in the coding region for the tRNA for glutamic acid (fig. 1). The first eight bases at the 3' end of primer MVZ 05 are totally conserved in the human (Anderson et al. 1981), bovine (Anderson et al. 1982), mouse (Bibb et al. 1981), and rat (Gadaleta et al. 1989) sequences, suggesting that this region is well conserved in mammals. Primer MVZ 05 (=L14724 in Irwin et al. 1991) has worked well in more than 20 species of mammals (Irwin et al. 1991; D. Irwin, personal communication). Presumably, primer MVZ 05 makes a good match with the tRNA adjacent to the functional copy of the cyt b gene in *C. jelskii*. As a test of the mismatch hypothesis, primer MVZ 03 was replaced by MVZ 07, a primer designed specifically to match *Akodon* sequences. With this new primer combination (MVZ 07—MVZ 04) the functional sequence was obtained from the individuals of *C. jelskii*.

Comparison of the Two Forms of cyt b, and Evidence in Favor of a Nuclear Copy

The sequence of the deletion copy of cyt b in *C. jelskii* differs greatly from the
Primer MVZ 03  
CCA TCC AAC ATC TCA GCA TGA TG-3'

Chroeomys jelskii  
..C .G. ..T ..T ..T T.T .......

Bolomys amoenus  
... .T ... .T ... T.C ......

Akodon boliviensis  
... .T ...... ... .T.C ......

*Fig. 4.*—Sequence data obtained using primer pair MVZ 05-MVZ 04, showing number of mismatches with primer MVZ 03 in *Chroeomys jelskii* compared with two other representative species of akodontine rodents (also see Smith and Patton 1991).

functional sequence. For the 285 sites in common where the bases are known, comparisons from the two geographic samples of *C. jelskii* have 71 or 75 nucleotide differences, or 25%-26% sequence difference (table 1). The mtDNA sequence in *Mus domesticus* differs from that of the nuclear copies in *Chroeomys* by 74 or 76 bp. On the other hand, the mtDNA sequence in *Mus* differs from the mtDNA sequence in *Chroeomys* by only 71 or 72 differences. Thus, the deletion sequence in *Chroeomys* differs from the functional one by at least as many bases as *Mus* differs from *Chroeomys*. Comparison of the mitochondrial sequence in *M. domesticus* versus the nuclear sequences shows that 47% of the changes cause replacements. Comparisons of the mitochondrial sequences in *Chroeomys* versus the nuclear sequences show 32%-35% of the changes as replacements. Finally, comparisons among the mitochondrial sequences show the lowest number of replacement changes (14.3%-26.4%).

The nonfunctional copy could be located in the mitochondrial genome either as an intramitochondrial duplication (for examples of duplications involving cyt b in several species of lizards, see Moritz and Brown 1987) or in the form of heteroplasmy, with two mtDNA haplotypes within each individual (cases of heteroplasmy with a deletion in one form are described in Boursot et al. 1987; Holt et al. 1988; Dowling et al. 1990). Alternatively, the anomalous copy might be in the nucleus. Examples of nuclear sequences that are homologous with portions of the mitochondrial genome have been described in eight organisms (for references, see Introduction).

The estimated total size of the *C. jelskii* mitochondrial genome, on the basis of *EcoRI*, *AvaII*, and *HinfI* digests (fig. 5 and data not shown), is 15 or 16 kb. This mtDNA genome size in *C. jelskii* is typical of mammalian genomes and is not consistent with any large duplications, although a small duplication cannot be ruled out.

The restriction digests showed only one set of bands, ruling out the possibility of significant amounts of two forms of the mitochondrial gene in one individual (fig. 5). Heteroplasmy also seems unlikely in the case of *C. jelskii*, because the two heteroplasmic forms would have to have been maintained for a long period of evolutionary time in order to reach such extreme differentiation, or else there would have to have been a paternal contribution.

When purified mtDNA from two individuals of *C. jelskii* (MVZ 173073 and MVZ 173083) was used as the template, the primer pair MVZ 03–MVZ 04 produced double-stranded product of the same size as the normal fragment in other species in the *Akodon* group (fig. 6). The normal size fragment was consistently obtained in experiments with annealing temperatures of 45°C, 50°C, and 55°C. In extracts of total cellular DNA the primer pair MVZ 03–MVZ 04 amplifies the nonfunctional nuclear copy of cyt b under annealing temperatures as high as 60°C, but in purified mtDNA from the same individuals the primer pair amplifies only the functional mitochondrial gene.
Table 1
Sequence Differences among Mitochondrial, (Functional) and Nuclear (Nonfunctional) Copies of Cytochrome b

<table>
<thead>
<tr>
<th>GENE</th>
<th>MITOCHONDRIAL</th>
<th>NUCLEAR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MVZ 173074</td>
<td>MVZ 173083</td>
</tr>
<tr>
<td>Mus mitochondrial</td>
<td>36/35</td>
<td>36/36</td>
</tr>
<tr>
<td>173074 mitochondrial</td>
<td>51/20</td>
<td>13/1</td>
</tr>
<tr>
<td>173083 mitochondrial</td>
<td>53/19</td>
<td>12/2</td>
</tr>
<tr>
<td>173074 nuclear</td>
<td>40/36</td>
<td>50/25</td>
</tr>
<tr>
<td>173083 nuclear</td>
<td>39/35</td>
<td>51/24</td>
</tr>
</tbody>
</table>

NOTE.—Data are for pairwise comparisons among Mus and the mitochondrial (functional) and nuclear (deletion) copies of cyt b in a representative individual of Chirogaleus jelskii from each of the two geographic localities. Values above the diagonal are the number of transitions/transversions; below the diagonal are the number of silent/replacement changes. The comparisons are based on the 285 sites having no gaps (X's in fig. 3).

An additional experiment was performed to test whether both the functional and nonfunctional sequences were present in the mitochondrial genome in C. jelskii. Purified mtDNA from C. jelskii was digested with MboI, which cuts twice in the functional sequence (at positions 14196 and 14441, numbered relative to the M. domesticus sequence published in Bibb et al. 1981) but only once (at position 14348) in the deletion sequence. When purified C. jelskii mtDNA was probed with the functional copy of the cyt b sequence, a band of the appropriate size (245 bp) was observed (fig. 7). When the same digest was probed with the deletion copy, the same 245-bp band was observed. If a single mitochondrial genome contained both functional and nonfunctional copies of cyt b, then the deletion probe should have hybridized to two different fragments of sizes different than that observed. The minimum size of these nonfunctional fragments would be 110 and 255 bp if additional MboI sites were located immediately outside the primers.

Mitochondrial sequences apparently move relatively readily into the nucleus. Fukuda et al. (1985) suggest that the integration of mtDNA fragments into nuclear DNA is mediated by mechanisms similar to those seen in the case of the integration of viral DNA into host nuclear DNAs. Thorsness and Fox (1990) found that, in yeast, sequences from mitochondria moved into the nucleus 100,000 times more frequently than nuclear sequences moved into the mitochondria.

If the anomalous sequence in C. jelskii is translated using the nuclear genetic code for amino acids, there are still several termination codons, so this sequence cannot produce a functional protein in the nucleus. The 25%–26% sequence difference for the mitochondrial versus nuclear copies in C. jelskii is somewhat less than that which Fukuda et al. (1985, fig. 6) describe for the most distant pairs of mitochondrial versus nuclear copies in humans. The Lm H-4 clone studied by Fukuda et al. (1985) differs from the human mtDNA sequence by 32.6% at ND5, and the Lm P-6 clone differs from the human mtDNA sequence by 31.4% at ND5.

Rates of Change

It is possible that other species of akodontines—and even more distantly related rodents—could also have this mtDNA-like sequence in their nucleus. As detailed in
fig. 5.—agarose gel of EcoRI restriction-endonuclease digest of purified mtDNA from Chroeomys jelskii (MVZ 173083) (lane 1). the gel was 2% NuSieve in 1× TA buffer (Maniatis et al. 1982). the size standard is a HindIII digest of Lambda (lane 2).

Table 1, the Mus and Chroeomys mtDNA sequences are as similar to each other as either is to the nuclear copy, on the basis of total number of base-pair differences. This suggests that the nuclear invasion may have occurred around the time of the split between Mus and the South American akodontine rodents. However, if the mitochondrial copies in Mus and Chroeomys are evolving at a much higher rate and under different constraints than is the nuclear copy, the relative depth of the Mus lineage (fig. 8) may be due to its rapid rate of evolution, making it possible that the nuclear copy arose prior to the Mus-akodontine split.

Typical of all mammalian mitochondrial genomes, the sense strand of the Chroeomys cyt b gene is extremely biased against G at the third positions of codons (5 G’s in 100 codons) (see Brown 1985). Given the extent of divergence between the nuclear and mitochondrial copies, it is interesting that the proposed nuclear non-functional copy is still strongly biased against G at third positions (6 in 94 codons). One explanation for the apparent maintenance of this strong bias could be that most of the evolution since the transfer of the mitochondrial copy to the nucleus has occurred in the mitochondrial lineage.
The nuclear copies in the two localities, as represented by individuals MVZ 173074 and MVZ 173083 (fig. 8), have accumulated only two differences in the 285 bp compared. On the other hand, 17 substitutions are inferred on the tree (fig. 8) for the mitochondrial sequences, as represented by the same two individuals. It is possible that the evolutionary history of the nuclear and mitochondrial genes is different and that the divergence of the mitochondrial sequences occurred much earlier than did that of the nuclear copies we have sampled. However, the pattern we have observed is consistent with the suggestion that the rate of change in mtDNA is indeed faster than that in nuclear DNA—and that therefore the large difference between the mitochondrial and nuclear copies is due mainly to the high rate of change in the protein-coding mitochondrial gene.

Precautions for PCR

The use of versatile primers (Kocher et al. 1989) has been an extremely productive approach, allowing researchers to obtain and compare sequences from many different taxa. However, our results highlight the need to be alert to the fact that targets other than the one of interest can be amplified. In the above analysis we present an example of an aberrant amplification of part of the cyt b gene. The nonfunctional copy is probably located in the nuclear genome, evolving under different constraints, and would present many potential problems if included in an evolutionary analysis. In this example our putative nuclear copy of the cyt b
Fig. 7.—Southern hybridization of 20 ng of purified mtDNA (MVZ 173083) digested with *Mbo*I, separated on a 4% NuSieve TBE agarose gel, and transferred to a nitrocellulose filter (Maniatis et al. 1982, pp. 382–389). Lane 1, mtDNA probed with functional copy amplified using primer pair MVZ 05–MVZ 04. Lane 2, mtDNA probed with deletion copy amplified with primer pair MVZ 03–MVZ 04 from liver DNA.

gene first raised suspicions because (a) the double-stranded product was smaller than that in other species, (b) the sequence had a deletion, and (c) the reading frame contained several stop codons. However, a nuclear copy would not always have a deletion that would throw off the reading frame and flag the aberrant sequence. Furthermore, non-protein-coding sequences such as rDNA and noncoding regions of the mitochondrial genome are also employed in assessing variation in

Fig. 8.—Evolutionary tree showing most parsimonious branching order for mitochondrial and nuclear copies of cyt b sequences from *Chromomys jelskii* individuals. The *Mus domesticus* cyt b sequence (Bibb et al. 1981) is used as an outgroup. The branching order is based on the sequences shown in fig. 3. The number of inferred substitutions is shown next to each branch.
populations and in phylogenetic studies. For these noncoding regions the criterion of the maintenance of the reading frame cannot be used to verify that the sequence obtained is the one of interest.

In cases where a protein-coding mitochondrial sequence has been transferred to the nucleus, the determination of functionality is useful (see also Irwin et al. 1991)—but only for comparisons where the transfer of one copy to the nucleus is a relatively ancient event, which would allow for substantial evolutionary change. For more recent transfer events, where there is little accumulated variation, one way to check that the sequence being amplified is not from the nucleus would be to use purified mtDNA. However, there are still circumstances where this would not be a definitive test. For example, if there were a rearrangement in the mitochondrial genome, then the primers might not be able to amplify a mitochondrial copy and might instead amplify a nuclear copy from the very small amount of nuclear material in the purified mitochondrial extract. A requirement to work from purified mtDNA would negate the tremendous advantage of being able to obtain sequence from preserved material other than frozen tissues. Investigators will have to weigh these concerns for each particular project.

An even more difficult problem could arise if the aberrant copy is located in the mitochondrial genome as either a duplication (on the same genome) or in the heteroplasmic state (two different genomes in the same individual). In such cases the events will often be recent and only detectable by detailed analysis of the mitochondrial genome organization.

Acknowledgments

Monica Frelow, Cristian Orrego, and Tom White provided advice and assistance with PCR and sequencing. We had helpful discussions with Tom Kocher. David Irwin, Craig Moritz, Tom Quinn, and Walter Fitch provided helpful comments on the manuscript. Karen Klitz assisted with several of the figures. Financial support was provided by the Museum of Vertebrate Zoology and by grants from the NSF to J.L.P. and by a National Institutes of Health postdoctoral fellowship to W.K.T.

LITERATURE CITED


WESLEY M. BROWN, reviewing editor

Received June 27, 1991; revision received September 6, 1991

Accepted September 6, 1991
Coevolution of the Vertebrate Integrin α- and β-Chain Genes

Austin L. Hughes
Department of Biology and Institute of Molecular Evolutionary Genetics, Pennsylvania State University

The integrin receptors are heterodimers whose α and β subunits are encoded by separate, evolutionarily unrelated multigene families. Phylogenetic analysis of DNA sequences from these two gene families showed that they have not always evolved in a parallel fashion. The integrin α chains that can form heterodimers with β1 do not constitute a monophyletic group, nor do the β chains which can form heterodimers with αV. On the other hand, the vertebrate α chains associating with β2 are a monophyletic group. In the metal cation binding region of the α chain, an exon exchange took place between human αM and αX ~40–50 Mya, homogenizing this functionally important region in these two α chains. When integrin β chains of different functional classes are compared, nonsynonymous (amino acid altering) nucleotide substitutions that alter amino acid residue charge in the central region of the molecule occur at a rate significantly higher than that expected under random replacement. By contrast, when closely related β1 chains are compared, residue charge is conserved in this region. These results pinpoint the central region as a focus of functional divergence among integrin β chains, perhaps relating to the ability of each β integrin class to associate with a specific array of α integrins. Furthermore, they imply that positive, directional selection on this region has occurred in the evolution of the integrin β-chain gene family.

Introduction

In eukaryotes many proteins are heterodimers or heterooligomers that are made up of evolutionarily unrelated or distantly related subunits. In some cases, there exists a family of such heterodimers, whose members perform distinct but related functions; and the subunits of such a family of heterodimers are encoded by separate yet coevolving gene families. How such coevolution occurs between functionally interacting gene families is not well understood. Here I examine this process in the case of integrin α- and β-chain gene families of vertebrates, which encode the α and β subunits of the integrin family of receptors.

Integrins are cell-surface glycoproteins that function in cell-cell and cell-matrix adhesion. In vertebrates, they are involved in such fundamental physiological processes as leukocyte adhesion at the site of inflammation, binding of cytotoxic T-cells to target cells, binding of macrophages to complement C3bi, adhesion of activated platelets, and development of muscle (Neff et al. 1982; Hynes 1987; Ruoslahti and Pierschbacher 1987; Larson and Springer 1990; Springer 1990). In Drosophila, they play an essential role in development of muscles and in the attachment of mesoderm to ectoderm.
The integrin α and β subunits associate non-covalently; and the two gene families that encode them are evolutionarily unrelated. Each α chain can associate with one or more β chains to form an array of receptors; likewise, each β chain can associate with one or more α chains. However, only a small fraction of all conceivable α/β heterodimers actually is found in vivo. Here I estimate the phylogenetic relationships of integrin α- and β-chain genes in order to examine how phylogeny correlates with the ability of specific α and β chains to form functional heterodimers.

The process by which members of a multigene family diverge and assume new functions is fundamental to adaptive evolution (Li 1983). Two major mechanisms are thought to be involved in this process: (1) recombination, either within a gene family or between unrelated genes, may give rise to a novel protein capable of a new function (Gilbert 1978; Doolittle 1985); and (2) directional positive selection favoring certain amino acid replacements in functionally important protein regions may gradually lead to functional divergence between related proteins (Waley 1969; Jensen 1976). Here I test for evidence of both of these mechanisms in the history of integrin α- and β-chain genes. I test for evidence of the former mechanism by constructing separate phylogenies for different gene regions. Previously it has proved difficult to obtain evidence for the latter mechanism because absence of conservation of amino acid sequence in a particular region between two related proteins may indicate either absence of a functional constraint on this region in either protein or functional divergence between the two proteins in this region. Here a method of distinguishing between these two possibilities (Hughes et al. 1990) is applied to analysis of integrin α and β chains.

**DNA Sequences Analyzed**

Table 1 lists integrin α- and β-chain gene sequences analyzed in the present paper, with information both on the heterodimers formed by each and on the function of some of the better-known integrin receptors. The majority of available sequences are from the human. For αM and β2, both human and mouse sequences are available. β1 sequences are available for human, mouse, chicken, and the frog *Xenopus laevis*. In the case of *X. laevis*, there are two functional β1 chains, β1a and β1b, for both of which DNA sequence is available. This duplication of β1 integrins is believed to be a result of the tetraploidization of *Xenopus*, which took place ~40 Mya (DeSimone and Hynes 1988).

The structure of the integrin receptor is not known in detail, but certain functionally important regions have been identified (fig. 1). In certain α chains, there is an inserted domain (I-DOM; 166–190 amino acids, depending on the individual α chain) that is homologous to domains in cartilage matrix protein, von Willebrand factor, complement component C2, and factor B (Corbi et al. 1988; Takada and Hemler 1989). All α integrins have a domain consisting of seven repeated units of ~50–60 amino acids each. This region contains putative sites for binding metal cations (Ca\(^{2+}\) or Mg\(^{2+}\)). There are three such cation-binding sites in those integrins having an I-DOM; these three binding sites are found in the three terminally located repeat units. In other α chains, there are four cation-binding sites, located in the four C-terminally located repeat units. Putative transmembrane regions (TM) have also been identified in α chains, but, unlike the case of β chains (Suzuki and Naitoh 1990), the putative TM of a given vertebrate α chain does not necessarily align precisely with those of other vertebrate α chains (Larson et al. 1989; Takada and Hemler 1989).
<table>
<thead>
<tr>
<th>Name</th>
<th>Species (reference)*</th>
<th>Associated with</th>
<th>Name(s)</th>
<th>Receptors</th>
<th>Ligands</th>
</tr>
</thead>
<tbody>
<tr>
<td>α1b</td>
<td>R (1)</td>
<td>β1</td>
<td>VLA-1 (α1β1)</td>
<td>Laminin and collagen</td>
<td></td>
</tr>
<tr>
<td>α2b</td>
<td>H (2)</td>
<td>β1</td>
<td>VLA-2 (α2β1)</td>
<td>Laminin and collagen</td>
<td></td>
</tr>
<tr>
<td>α4</td>
<td>H (3)</td>
<td>β1 and βp</td>
<td>VLA-4 (α4β1)</td>
<td>VCAM-1 and fibronectin</td>
<td></td>
</tr>
<tr>
<td>α5</td>
<td>H (4)</td>
<td>β1</td>
<td>VLA-5 (α5β1)</td>
<td>Fibronectin&lt;sup&gt;‡&lt;/sup&gt; (RGD)</td>
<td></td>
</tr>
<tr>
<td>αLb</td>
<td>H (5)</td>
<td>β2</td>
<td>LFA-1 (αLβ2)</td>
<td>ICAM-1,2</td>
<td></td>
</tr>
<tr>
<td>αMb</td>
<td>H (6) and M (7)</td>
<td>β2</td>
<td>Mac-1 (αMβ2)</td>
<td>C&lt;sup&gt;3&lt;/sup&gt;bi, factor X, and fibrinogen</td>
<td></td>
</tr>
<tr>
<td>αXb</td>
<td>H (8)</td>
<td>β2</td>
<td>p150,95 (αXβ2)</td>
<td>Unknown</td>
<td></td>
</tr>
<tr>
<td>αIib</td>
<td>H (9)</td>
<td>β3</td>
<td>gpiⅧ/Ⅲa (αIibβ3)</td>
<td>Fibronectin and fibrinogen&lt;sup&gt;‡&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>αV</td>
<td>H (10)</td>
<td>β1, β3, and β5</td>
<td>CD51/CD29 (αVβ1)</td>
<td>Fibronectin, fibrinogen, and von Willebrand factor&lt;sup&gt;‡&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>αPS2</td>
<td>D (11)</td>
<td>βPS</td>
<td>VNR (αVβ3)</td>
<td>Vitronectin, fibrinogen, von Willebrand factor, and thrombospondin&lt;sup&gt;‡&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>β1</td>
<td>C (12)</td>
<td>α1, α2, and α3</td>
<td>CD51/CD2- (αVβ5)</td>
<td>Vitronectin and fibronectin&lt;sup&gt;‡&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>H (4)</td>
<td>α4, α5, and α6</td>
<td>αPS2βPS</td>
<td>Unknown</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M (13)</td>
<td>αV</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>X (14)</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>β2</td>
<td>H (15)</td>
<td>αL, αM, and αX</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>M (16)</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>β3</td>
<td>H (17)</td>
<td>αIib and αV</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>β4</td>
<td>H (18)</td>
<td>α6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>β5</td>
<td>H (19)</td>
<td>αV</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>β6</td>
<td>H (20)</td>
<td>?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>βPS</td>
<td>D (21)</td>
<td>αPS1 and αPS2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note.—Functional information is from Springer (1990) and Volk et al. (1990).

* C = chicken; D = Drosophila melanogaster; H = human; M = mouse; R = rat; and X = Xenopus laevis. (1) = Ignatius et al. (1990); (2) = Takada and Hemler (1989); (3) = Takada et al. (1989); (4) = Argraves et al. (1987); (5) = Larson et al. (1989); (6) = Corbi et al. (1988); (7) = Pytel et al. (1988); (8) = Corbi et al. (1987); (9) = Ponce et al. (1987); (10) = Suzuki et al. (1987); (11) = Bogaert et al. (1987); (12) = Tamkun et al. (1986); (13) = Tominaga (1988); (14) = DeSimone and Hynes (1988); (15) = Kishimoto et al. (1987); (16) = Wilson et al. (1989); (17) = Fitzgerald et al. (1987); (18) = Suzuki and Naitoh (1990); (19) = Ramaswamy and Hemler (1990); (20) = Sheppard et al. (1980); (21) = Mackrell et al. (1988).

<sup>‡</sup> Has I-DOM.

<sup>‡</sup> Known to bind RGD motif.
Certain integrin receptors crosslink to ligand domains containing the sequence arginine-glycine-aspartic acid (RGD) (table 1). In the case of the β3 subunit, the receptor for this crosslinking is located at residues 109–171 of the mature protein; this is a highly conserved region in integrin β chains, even in those not known to have an RGD crosslinking function (D’Souza et al. 1988; Springer 1990). The α chain may also be involved in RGD crosslinking through an interaction among RGD ligand motif, the β and α chains, and the metal cation (Corbi et al. 1987; Loftus et al. 1990).

In all integrin β chains, there is a region rich in conserved asparagine residues, several of which have been proposed as potential N-glycosylation sites. There is a cysteine-rich region consisting of four repeats of a 40–50-amino-acid segment with about eight cysteines per repeat; in this region extensive intrachain disulfide bonding is believed to occur (Hynes 1987). TM consists of the same number of residues in all known integrin β chains. In β4 the cytoplasmic portion (CYT) is much longer (~1,000 residues) than in other β integrins (~50 residues). The elongated CYT of β4 is believed to be homologous to a repeating region of fibronectin and twitchin (Suzuki and Naitoh 1990).

In this research integrin α and β amino acid sequences were aligned by the method of Gotoh (1987) and were corrected by eye in some instances, to conform with published alignments. (These alignments are available on request.) In analysis of DNA sequences, any codon where the alignment postulated a gap (deletion/insertion) was omitted from all comparisons; thus a comparable data set was used in each pairwise comparison.

To test for domain exchange and to examine evolutionary forces acting on different protein regions, I analyzed different regions separately. The correspondence between exons and protein domains is known only for human αX (Corbi et al. 1990), because for other α- and β-chain genes only cDNA sequences have been published. For the regions analyzed, see figure 1. Only the conserved metal-binding region (MBR) of the

---

**Fig. 1.—Schematic illustration of integrin α- and β-chain structure.** In the case of α chains, the following regions were analyzed: (1) NTR of the mature protein, exclusive of the I-DOM (214 codons); (2) I-DOM (166 codons); (3) MBR, including the three cation-binding sites common to all α integrins, a region corresponding to exons 14 and 15 of human αX (144 codons); and (4) CTR, including the remainder of the extracellular portion of the molecule and the transmembrane portion (470 codons). CYT of α chains was not included in analyses, because only a small number of residues could be aligned between distantly related molecules. For the purpose of some analyses, the CTR was divided into two regions: CEN (441 codons) and TM (29 codons). In the case of the integrin β chains, the following regions were analyzed: (1) NTR of the mature protein (91 codons); (2) the region homologous to the RGD ARE (61 codons); (3) the NRR (87 codons); (4) the CEN, excluding NRR but including cysteine-rich portions (341 codons); (5) the premembrane region (PM) (67 codons); (6) TM (23 codons); and (6) CYT (46 codons).
αPS2 gene of *Drosophila melanogaster* was aligned with other α chains, because sequence similarity outside this region was very low. Similarly, for both the *Drosophila* βPS and the human β4, which are highly divergent from other known β chains, only three conserved regions—the adhesion receptor region (ARE), the asparagine-rich region (NRR), and TM—were aligned with other β chains.

**Results**

**Phylogenetic Relationships of β Chains**

Figure 2 shows a phylogenetic tree of integrin β chains that is based on number of nonsynonymous nucleotide substitutions per site ($d_N$) (Nei and Gojobori 1986) in ARE, NRR, and TM. $d_N$ was used because in many comparisons synonymous sites are saturated, so that the number of synonymous substitutions per site ($d_S$) cannot be computed. The β-chain genes form two major clusters: (1) β1/β2 and (2) β3/β5/β6. The *Drosophila* βPS clustered with β1/β2 rather than with β3/β5/β6, but this clustering was not statistically significant. Because amphibian, bird, and mammal β1 integrins cluster together (fig. 2), it seems likely that β1 diverged from other integrin β chains prior to the divergence of amniotes and amphibians (~370 Mya); however, sequences of other β chains from amphibiaans are needed to confirm this hypothesis.

Phylogenetic trees of the vertebrate integrin β chains, based either on the entire

![Phylogenetic Tree](image)

**Fig. 2.—** Phylogenetic tree of integrin β chains, based on $d_N$ in ARE, NRR, and TM. The tree was constructed by the neighbor-joining method (Saitou and Nei 1987). Since β4 is the most divergent known β chain, it was chosen as an outgroup to root the tree. Prefixes indicate species as follows: C = chicken; D = *Drosophila melanogaster*; H = human; M = mouse; X = *Xenopus laevis*. The double asterisks (**) denote the branch length that is significantly different from zero at the 1% level (Li 1989).
coding region or on different regions of the genes, did not differ significantly in topology from the tree shown in figure 2. Thus, there was no evidence of interlocus recombination among the integrin β-chain genes.

β3 and β5 are closely related both phylogenetically (fig. 2) and functionally. Both associate with αV and form receptors that bind RGD (table 1). On the other hand, β1 and β2, which cluster together in the phylogenetic tree, share no close functional relationship. β2 associates with a set of α chains (αL, αM, and αX) which associate with no other β chain, whereas β1 associates with a large number of α chains, one of which (αV) associates with all other β chains except β2.

Phylogenetic Relationships of α Chains

Phylogenetic trees were constructed separately for the N-terminal region (NTR), I-DOM, MBR, and the C-terminal region (CTR) of vertebrate α chains (fig. 3). In the tree for MBR, αPS2 was also included. Trees were constructed by the neighbor-joining method based on dN. In each gene region, the α chains having I-DOM clustered together (fig. 3); thus, the results suggest that the exchange of I-DOM to α integrins was a unique event. The αL, αM, and αX chains, all of which associate with β2, form a closely related group in all trees (fig. 3). Likewise, α5, αIIb, and αV, which form
heterodimers involved in RGD crosslinking, form a cluster. In MBR, the *Drosophila* αPS2 gene is a member of this cluster [fig. 3(C)]. On the other hand, the α chains that associate with β1 (α1, α2, α4, α5, and αV) are a very disparate group phylogenetically. Among them only α1 and α2 are closely related.

The trees based on NTR [fig. 3(A)] and CTR [fig. 3(D)] have identical topologies. For α integrins having I-DOM, the tree based on I-DOM [fig. 3(B)] has the same topology as do the NTR and CTR trees. However, in MBR [fig. 2(C)], there is a statistically significant departure from the topology of the other trees. In NTR, I-DOM, and CTR, human αM and mouse αM genes cluster together; but in MBR human αM clusters with human αX. The clustering of human αM and αX in MBR is not a result of the presence of αPS2 in the MBR tree; the same clustering was also seen when αPS2 was excluded from the analysis (data not shown). The MBR tree also differs from the trees for NTR and CTR, in that α5 clusters with αIb in the former and with αV in the latter; however, this difference is not statistically significant (fig. 3).

The similarity of human αM and αX in MBR suggests that an interlocus exchange of MBR may have occurred in the history of the two genes. Further analyses were conducted to decide between this and alternative hypotheses. Unexpectedly high similarity between two genes at nonsynonymous sites in a particular region might be due to either convergent evolution or shared constraint at the amino acid level. By this hypothesis, a similar pattern would not be expected to occur at synonymous sites (Hughes 1991 b). On the other hand, a shared codon-usage bias might produce synonymous-sites similarity between two distantly related genes (Wolfe et al. 1989).

In order to decide among these hypotheses, $d_s$ and $d_N$ were computed among human αM, mouse αM, and human αX in different gene regions (table 2). In MBR, both $d_s$ and $d_N$ between human αM and human αX are significantly lower than those between human and mouse αM; and both $d_s$ and $d_N$ between human αM and αX are significantly lower in MBR than in any other region (table 2). In regions other than MBR, $d_N$ between human αM and αX is always significantly higher than that between human and mouse αM; and $d_s$ between human αM and αX in these regions is also lower than that between human and mouse αM, although the difference is significant only in I-DOM (table 2).

To test whether a shared biased G+C content at third-codon positions (Wolfe et al. 1989) can account for the relatively low $d_s$ between human αM and αX in MBR, percentage of G+C at third-codon positions and at all positions was calculated for different gene regions (table 3). Integrin α-chain genes were found to vary greatly in third-position G+C content. Third-position G+C content is <50% in all regions of α2, α4, and αV. On the other hand, in rat α1 it is close to 50% in all regions; and in other α-chain genes it is >50% in all regions and is >60% in most regions. It is interesting that, even though I-DOM was exchanged to the integrin α chains by a single event, the third-position G+C content of I-DOM now resembles that of the rest of the gene, being low in α2, intermediate in α1, and high in other α integrins with I-DOM (table 3).

Human αM and αX share a high third-position G+C content in MBR, whereas G+C content of mouse αM in MBR is somewhat lower than that of either of these two human genes (table 3). However, the shared G+C-content bias of human αM and αX in MBR cannot explain the low $d_s$ between them in this region. Human αM and αX share a high third-position G+C content in other regions as well, but $d_s$ between them is much higher in other regions than it is in MBR (tables 2 and 3).

Therefore, the data support the hypothesis that in the human lineage, sometime after the divergence of mouse and human, MBR of αM was donated to αX. At both
Table 2

*d* ± Standard Error (SE) per 100 Synonymous Sites *d*_ ± SE per 100 Nonsynonymous Sites, in Comparisons of Human and Mouse αM and Human αX Genes

<table>
<thead>
<tr>
<th>REGION</th>
<th>HUMAN αM vs. MOUSE αM</th>
<th>HUMAN αM vs. HUMAN αX*</th>
<th>MOUSE αM vs. HUMAN αX*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>d</em>_ ± <em>dN</em></td>
<td><em>d</em>_ ± <em>dN</em></td>
<td><em>d</em>_ ± <em>dN</em></td>
</tr>
<tr>
<td>NTR</td>
<td>62.7 ± 9.3</td>
<td>13.9 ± 1.8</td>
<td>75.7 ± 11.1***</td>
</tr>
<tr>
<td>I-DOM</td>
<td>50.4 ± 8.9</td>
<td>11.7 ± 1.9</td>
<td>226.1 ± 87.3*</td>
</tr>
<tr>
<td>MBR</td>
<td>60.7 ± 10.3</td>
<td>11.6 ± 2.0</td>
<td>30.1 ± 6.0*</td>
</tr>
<tr>
<td>CTR</td>
<td>76.7 ± 7.8</td>
<td>16.1 ± 1.4</td>
<td>96.3 ± 10.0***</td>
</tr>
<tr>
<td>All</td>
<td>69.6 ± 5.2</td>
<td>14.7 ± 1.0</td>
<td>73.2 ± 5.4</td>
</tr>
</tbody>
</table>

Tests of significance of difference between *d*_ or *dN* and corresponding value for same region in human αM vs. mouse αM comparison: * = *P* < .05; ** = *P* < .01; and *** = *P* < .001. Tests of significance of difference between *d*_ or *dN* and corresponding value for MBR: † = *P* < .05; †† = *P* < .01; and ††† = *P* < .001.
Table 3
% G+C Content at Third-codon Positions (3d) and at All Positions in Regions of Integrin α-Chain Genes

<table>
<thead>
<tr>
<th>GENE</th>
<th>NTR</th>
<th></th>
<th>I-DOM</th>
<th></th>
<th>MBR</th>
<th></th>
<th>CTR</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3d</td>
<td>All</td>
<td>3d</td>
<td>All</td>
<td>3d</td>
<td>All</td>
<td>3d</td>
<td>All</td>
</tr>
<tr>
<td>αM</td>
<td>68.7</td>
<td>55.9</td>
<td>62.7</td>
<td>49.6</td>
<td>75.0</td>
<td>65.0</td>
<td>74.8</td>
<td>55.4</td>
</tr>
<tr>
<td>M-αM</td>
<td>61.2</td>
<td>51.4</td>
<td>59.6</td>
<td>48.6</td>
<td>59.7</td>
<td>57.9</td>
<td>62.6</td>
<td>50.3</td>
</tr>
<tr>
<td>αX</td>
<td>79.9</td>
<td>60.9</td>
<td>58.4</td>
<td>47.2</td>
<td>79.2</td>
<td>66.2</td>
<td>75.2</td>
<td>57.2</td>
</tr>
<tr>
<td>αL</td>
<td>65.4</td>
<td>58.4</td>
<td>63.3</td>
<td>47.6</td>
<td>63.9</td>
<td>59.7</td>
<td>71.0</td>
<td>55.2</td>
</tr>
<tr>
<td>α2</td>
<td>37.4</td>
<td>45.5</td>
<td>30.1</td>
<td>39.4</td>
<td>43.1</td>
<td>44.9</td>
<td>40.1</td>
<td>41.3</td>
</tr>
<tr>
<td>R-α1</td>
<td>54.2</td>
<td>49.5</td>
<td>51.8</td>
<td>47.4</td>
<td>48.6</td>
<td>49.5</td>
<td>46.5</td>
<td>43.2</td>
</tr>
<tr>
<td>α4</td>
<td>37.7</td>
<td>45.6</td>
<td>...</td>
<td>...</td>
<td>34.7</td>
<td>45.1</td>
<td>37.4</td>
<td>40.2</td>
</tr>
<tr>
<td>α5</td>
<td>61.7</td>
<td>56.7</td>
<td>...</td>
<td>...</td>
<td>68.1</td>
<td>63.2</td>
<td>70.1</td>
<td>57.2</td>
</tr>
<tr>
<td>αV</td>
<td>40.2</td>
<td>46.0</td>
<td>...</td>
<td>...</td>
<td>34.0</td>
<td>45.6</td>
<td>39.9</td>
<td>41.6</td>
</tr>
<tr>
<td>αIIb</td>
<td>72.4</td>
<td>60.9</td>
<td>...</td>
<td>...</td>
<td>68.8</td>
<td>63.4</td>
<td>77.3</td>
<td>61.5</td>
</tr>
</tbody>
</table>

* M-αM = mouse αM. R-α1 = rat α1. Other genes are human.

synonymous and nonsynonymous sites, human αM and αX are about one-half as divergent from each other as either is from mouse αM (table 2). Therefore the recombinational event probably took place about one-half the time since the rodent/primate divergence, which is estimated to have taken place 80–100 Mya (Li et al. 1990). Because MBR corresponds to exons 14 and 15 of the αX gene and because αM presumably has similar genomic structure, an exon-shuffling mechanism is likely to have been involved.

The topology of the phylogenetic trees based on NTR and CTR (fig. 3A and C), which presumably portrays the true relationships among these genes, is identical to that obtained by Takada et al. (1989) for the seven of the nine α integrins analyzed here. Takada et al. applied the UPGMA method to percentage difference at the amino acid level; the UPGMA method depends on the assumption of a constant rate of evolution in each lineage, whereas the neighbor-joining method does not (Saitou and Nei 1987). The fact that these two methods produced the same result suggests that there are no marked differences in rate of evolution at nonsynonymous sites in different groups of vertebrate α integrins.

Adaptive Diversification of β Chains

When \( d_\alpha \) was computed in different regions of vertebrate integrin β-chain genes (table 4), ARE, NRR, and TM were found to be the most conserved regions. This was true both in comparisons among closely related β1 and β2 genes and in comparisons among the major groups of β-chain genes. In comparisons among different classes of integrin β chains, lack of conservation of the amino acid sequence of a particular region might indicate either a lack of constraint on the amino acid sequence of that region or adaptive diversification of that region in the different classes of integrin β chains. In order to discriminate between these two hypotheses, the method of Hughes et al. (1990) was applied to comparisons among integrin β-chain genes.

This method divides nonsynonymous nucleotide substitutions into those which are conservative and those which are radical (nonconservative) with respect to some amino acid residue side-chain property of interest. Similarly, each nonsynonymous nucleotide site is categorized as conservative or radical (or some fraction conservative
Table 4
Mean $d_N \pm SE$ per 100 Nonsynonymous Sites, in Comparisons of Different Regions of Integrin $\beta$-Chain Genes

<table>
<thead>
<tr>
<th>Region</th>
<th>X-(\beta)1a vs. H-(\beta)1</th>
<th>H-(\beta)1 vs. H-(\beta)2</th>
<th>$\beta$1 vs. (\beta)2 vs. (\beta)3/(\beta)5/(\beta)6 (N = 10)</th>
<th>(\beta)1 vs. (\beta)3/(\beta)5/(\beta)6 (N = 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTR</td>
<td>0.9 ± 0.7</td>
<td>4.6 ± 1.5*</td>
<td>11.3 ± 2.5**</td>
<td>59.1 ± 6.4***</td>
</tr>
<tr>
<td>ARE</td>
<td>0.0 ± 0.0</td>
<td>0.7 ± 0.7</td>
<td>2.2 ± 1.3</td>
<td>30.7 ± 5.6</td>
</tr>
<tr>
<td>NRR</td>
<td>1.0 ± 0.7</td>
<td>0.8 ± 0.6</td>
<td>2.0 ± 1.0</td>
<td>19.3 ± 3.2**</td>
</tr>
<tr>
<td>CEN</td>
<td>1.7 ± 0.0**</td>
<td>4.6 ± 0.8***</td>
<td>11.5 ± 1.2***</td>
<td>55.5 ± 3.2***</td>
</tr>
<tr>
<td>PM</td>
<td>0.0 ± 0.0</td>
<td>5.1 ± 1.2*</td>
<td>25.8 ± 5.7***</td>
<td>80.1 ± 9.0***</td>
</tr>
<tr>
<td>TM</td>
<td>2.1 ± 2.3</td>
<td>0.0 ± 0.0</td>
<td>10.1 ± 1.3</td>
<td>32.1 ± 1.1</td>
</tr>
<tr>
<td>CYT</td>
<td>0.9 ± 0.9</td>
<td>0.0 ± 0.0</td>
<td>1.8 ± 1.2</td>
<td>43.7 ± 7.9</td>
</tr>
<tr>
<td>All</td>
<td>0.8 ± 0.2</td>
<td>3.4 ± 0.5</td>
<td>10.0 ± 0.8</td>
<td>48.5 ± 2.0</td>
</tr>
</tbody>
</table>

Note.—$N =$ number of comparisons. SEs are estimated by Nei and Jin's (1989) method. For gene symbols, see table 1. Tests of significance of difference between $d_N$ and that for ARE: * $P < .05$; ** $P < .01$; and *** $P < .001$.

and some fraction radical), depending on the types of nonsynonymous change that can take place there. Both the proportion of conservative nonsynonymous differences per site ($P_{NC}$) and the proportion of radical nonsynonymous differences ($P_{NR}$) are then calculated. If $P_{NC} > P_{NR}$, then nonsynonymous substitutions have taken place in such a way as to conserve the amino acid property of interest. If $P_{NC} = P_{NR}$, then nonsynonymous substitutions have occurred at random, with respect to the property of interest. Thus, when there is no constraint with respect to a particular amino acid property, $P_{NC} = P_{NR}$. On the other hand, if $P_{NR} > P_{NC}$, then nonsynonymous nucleotide substitutions that change the amino acid property have occurred at a rate greater than expected under random substitution. Therefore, $P_{NR} > P_{NC}$ is evidence of directional positive selection leading to adaptive divergence between the two proteins.

In the case of integrin $\beta$ chains, $P_{NC}$ and $P_{NR}$ were computed with respect to both charge and polarity of residue side chains in different gene regions. With respect to charge, amino acids were divided into three categories: positive (H, K, and R), negative (D and E), and neutral (all others). With respect to polarity, there were two categories: nonpolar (A, F, I, L, M, P, W, and V) and polar (all others) (Lehninger 1975, pp. 73–75). Any nucleotide substitution leading to a change of category was counted as radical. The results of these analyses are shown in tables 5 (charge) and 6 (polarity).

In comparisons of closely related $\beta$1 chains (Xenopus laevis $\beta$1a vs. $\beta$1b and human vs. mouse $\beta$1), there was overall conservation of charge. This effect was more pronounced in the central extracellular portion (CEN) than in other variable regions; indeed, $P_{NC}$ with respect to charge significantly exceeded $P_{NR}$, in both of these comparisons (table 5). On the other hand, in comparisons among different classes of $\beta$ chains, charge was significantly conserved in the TM in all comparisons and in some cases was significantly conserved in ARE and CYT. Furthermore, in CEN all comparisons among different classes of $\beta$ chains showed a significant bias toward nonsynonymous substitutions causing a charge change (table 5). In all of these comparisons, a nonsynonymous difference was found at $\sim 50\%$ of the CEN sites where such a difference would cause a charge change (table 5).

The pattern of diversification with respect to polarity was quite different from that in the case of charge. Here there was an overall tendency to conserve polarity, both among closely related $\beta$1 and $\beta$2 chains and between major classes of $\beta$ chains.
Table 5

$\rho_{NC} \pm \text{SE}$ and $\rho_{NR} \pm \text{SE}$, with Respect to Residue Charge,
in Comparisons of Integrin $\beta$-Chain Genes

<table>
<thead>
<tr>
<th>Region</th>
<th>$\rho_{NC}$</th>
<th>$\rho_{NR}$</th>
<th>$\rho_{NC}$</th>
<th>$\rho_{NR}$</th>
<th>$\rho_{NC}$</th>
<th>$\rho_{NR}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTR</td>
<td>1.8 ± 1.2</td>
<td>0.0 ± 0.0</td>
<td>3.9 ± 1.8</td>
<td>5.0 ± 2.2</td>
<td>9.5 ± 2.7</td>
<td>11.7 ± 3.4</td>
</tr>
<tr>
<td>ARE</td>
<td>0.0 ± 0.0</td>
<td>0.0 ± 0.0</td>
<td>1.2 ± 1.0</td>
<td>0.0 ± 0.0</td>
<td>3.5 ± 1.7</td>
<td>0.0 ± 0.0*</td>
</tr>
<tr>
<td>NRR</td>
<td>0.8 ± 0.8</td>
<td>1.2 ± 1.2</td>
<td>5.7 ± 1.0</td>
<td>2.6 ± 0.9*</td>
<td>10.4 ± 1.4</td>
<td>11.2 ± 1.8</td>
</tr>
<tr>
<td>CEN</td>
<td>1.7 ± 0.6</td>
<td>0.0 ± 0.0**</td>
<td>8.2 ± 3.0</td>
<td>1.3 ± 1.3</td>
<td>23.2 ± 4.6</td>
<td>20.1 ± 4.8</td>
</tr>
<tr>
<td>TM</td>
<td>2.5 ± 2.5</td>
<td>0.0 ± 0.0</td>
<td>0.0 ± 0.0</td>
<td>0.0 ± 0.0</td>
<td>11.4 ± 5.1</td>
<td>0.0 ± 0.0*</td>
</tr>
<tr>
<td>CYT</td>
<td>1.8 ± 1.7</td>
<td>0.0 ± 0.0</td>
<td>0.0 ± 0.0</td>
<td>0.0 ± 0.0</td>
<td>3.5 ± 2.4</td>
<td>0.0 ± 0.0</td>
</tr>
<tr>
<td>All</td>
<td>1.3 ± 0.4</td>
<td>0.1 ± 0.1**</td>
<td>4.1 ± 0.6</td>
<td>2.2 ± 0.6*</td>
<td>9.4 ± 0.9</td>
<td>9.2 ± 1.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Region</th>
<th>$\rho_{NC}$</th>
<th>$\rho_{NR}$</th>
<th>$\rho_{NC}$</th>
<th>$\rho_{NR}$</th>
<th>$\rho_{NC}$</th>
<th>$\rho_{NR}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTR</td>
<td>39.5 ± 4.4</td>
<td>42.5 ± 4.8</td>
<td>43.6 ± 3.9</td>
<td>46.2 ± 4.5</td>
<td>47.1 ± 4.1</td>
<td>51.6 ± 5.1</td>
</tr>
<tr>
<td>ARE</td>
<td>23.2 ± 4.4</td>
<td>28.4 ± 6.0</td>
<td>29.4 ± 4.3</td>
<td>17.4 ± 3.9*</td>
<td>34.5 ± 4.7</td>
<td>32.0 ± 5.9</td>
</tr>
<tr>
<td>NRR</td>
<td>17.7 ± 3.3</td>
<td>16.1 ± 3.4</td>
<td>31.2 ± 3.6</td>
<td>21.7 ± 3.9</td>
<td>32.5 ± 3.8</td>
<td>22.2 ± 4.1</td>
</tr>
<tr>
<td>CEN</td>
<td>34.4 ± 2.1</td>
<td>46.1 ± 2.7***</td>
<td>37.9 ± 1.9</td>
<td>47.0 ± 2.6**</td>
<td>38.7 ± 2.0</td>
<td>52.6 ± 2.7***</td>
</tr>
<tr>
<td>PM</td>
<td>48.8 ± 5.2</td>
<td>49.7 ± 5.6</td>
<td>51.2 ± 4.7</td>
<td>52.7 ± 5.1</td>
<td>54.9 ± 4.8</td>
<td>58.9 ± 5.6</td>
</tr>
<tr>
<td>TM</td>
<td>31.6 ± 7.0</td>
<td>0.0 ± 0.0***</td>
<td>41.4 ± 6.8</td>
<td>5.8 ± 4.8***</td>
<td>50.8 ± 7.3</td>
<td>7.4 ± 7.5***</td>
</tr>
<tr>
<td>CYT</td>
<td>40.6 ± 6.5</td>
<td>25.2 ± 5.9</td>
<td>42.3 ± 5.6</td>
<td>26.3 ± 5.5*</td>
<td>52.7 ± 6.3</td>
<td>37.9 ± 6.6</td>
</tr>
<tr>
<td>All</td>
<td>33.5 ± 1.4</td>
<td>38.9 ± 1.8*</td>
<td>38.6 ± 1.3</td>
<td>40.1 ± 1.7</td>
<td>41.4 ± 1.4</td>
<td>46.0 ± 1.8*</td>
</tr>
</tbody>
</table>

Note.—All data are multiplied by 100. $N =$ number of comparisons. For gene symbols, see table 1. Tests of significance of difference between $\rho_{NC}$ and $\rho_{NR}$: $* = P < .05$; $** = P < .01$; and $*** = P < .001$.

In comparisons among different classes of $\beta$ chains, polarity was always significantly conserved in CEN. Thus, in comparisons among $\beta$-chain classes, CEN is characterized both by a significant bias toward amino acid replacements that lead to a charge change and by a significant tendency to conserve polarity. When the most cysteine-rich portion of CEN was analyzed separately, an identical pattern was seen both in this region and in the remainder of CEN (data not shown).

The bias that in CEN favors charge changes between different classes of integrin $\beta$ chains does not imply that these different classes differ markedly in overall proportions of positive, negative, and neutral amino acid residues. The largest difference in this respect is between $\beta 2$ and the other $\beta$ chains. For $\beta 2$, mean percentages of negative and positive residues in CEN are 12.8% and 10.9%, respectively; for $\beta 1$, they are 11.2% and 13.0%; and for $\beta 3$, $\beta 5$, and $\beta 6$, they are 11.1% and 12.2%. Thus it appears that it is the pattern of residue charges (charge profile) in CEN, rather than number of charged residues, that has been diversified among different classes of integrin $\beta$ chains.

In NTR and CYT, no directional bias in amino acid replacements was seen with respect to polarity or charge. Similar analyses with respect to amino acid residue volume and hydrophobicity likewise revealed no directional evolution in these regions (data not shown). Thus, it seems likely that the relatively low conservation in NTR and CYT is due to a relative lack of functional constraint. Previously it has been suggested that func-
Table 6
$p_{NC} \pm SE$ and $p_{NR} \pm SE$, with Respect to Residue Polarity, in Comparisons of Integrin $\beta$-Chain Genes

<table>
<thead>
<tr>
<th>REGION</th>
<th>$p_{NC}$</th>
<th>$p_{NR}$</th>
<th>$p_{NC}$</th>
<th>$p_{NR}$</th>
<th>$p_{NC}$</th>
<th>$p_{NR}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTR</td>
<td>1.4 ± 1.0</td>
<td>0.0 ± 0.0</td>
<td>6.0 ± 2.0</td>
<td>1.4 ± 1.4</td>
<td>14.3 ± 3.1</td>
<td>3.9 ± 2.2**</td>
</tr>
<tr>
<td>ARE</td>
<td>0.0 ± 0.0</td>
<td>0.0 ± 0.0</td>
<td>1.1 ± 1.0</td>
<td>0.0 ± 0.0</td>
<td>3.3 ± 1.9</td>
<td>0.0 ± 0.0</td>
</tr>
<tr>
<td>NRR</td>
<td>1.5 ± 1.0</td>
<td>0.0 ± 0.0</td>
<td>1.1 ± 0.9</td>
<td>0.0 ± 0.0</td>
<td>2.3 ± 1.3</td>
<td>1.4 ± 1.4</td>
</tr>
<tr>
<td>CEN</td>
<td>1.3 ± 0.5</td>
<td>0.4 ± 0.4</td>
<td>4.8 ± 0.9</td>
<td>3.6 ± 1.2</td>
<td>12.3 ± 1.4</td>
<td>7.3 ± 1.6</td>
</tr>
<tr>
<td>PM</td>
<td>0.0 ± 0.0</td>
<td>0.0 ± 0.0</td>
<td>7.1 ± 2.4</td>
<td>0.0 ± 0.0**</td>
<td>21.1 ± 4.0</td>
<td>23.2 ± 5.9</td>
</tr>
<tr>
<td>TM</td>
<td>3.7 ± 2.7</td>
<td>0.0 ± 0.0</td>
<td>0.0 ± 0.0</td>
<td>0.0 ± 0.0</td>
<td>16.9 ± 7.3</td>
<td>0.0 ± 0.0*</td>
</tr>
<tr>
<td>CYT</td>
<td>0.0 ± 0.0</td>
<td>2.9 ± 2.9</td>
<td>0.0 ± 0.0</td>
<td>0.0 ± 0.0</td>
<td>1.3 ± 1.3</td>
<td>2.9 ± 2.8</td>
</tr>
<tr>
<td>All</td>
<td>1.1 ± 0.3</td>
<td>0.4 ± 0.3</td>
<td>4.0 ± 0.6</td>
<td>1.8 ± 0.6**</td>
<td>10.8 ± 0.9</td>
<td>6.4 ± 1.0**</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>REGION</th>
<th>$p_{NC}$</th>
<th>$p_{NR}$</th>
<th>$p_{NC}$</th>
<th>$p_{NR}$</th>
<th>$p_{NC}$</th>
<th>$p_{NR}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTR</td>
<td>44.5 ± 4.0</td>
<td>34.2 ± 5.3</td>
<td>49.2 ± 3.5</td>
<td>36.3 ± 5.0</td>
<td>55.3 ± 4.1</td>
<td>37.9 ± 5.2**</td>
</tr>
<tr>
<td>ARE</td>
<td>26.7 ± 4.5</td>
<td>22.2 ± 5.9</td>
<td>29.4 ± 3.9</td>
<td>16.0 ± 4.3*</td>
<td>39.6 ± 4.8</td>
<td>21.6 ± 5.4*</td>
</tr>
<tr>
<td>NRR</td>
<td>17.8 ± 3.2</td>
<td>15.6 ± 4.2</td>
<td>29.7 ± 3.5</td>
<td>22.7 ± 4.3</td>
<td>30.9 ± 3.6</td>
<td>22.9 ± 4.3</td>
</tr>
<tr>
<td>CEN</td>
<td>45.4 ± 1.9</td>
<td>33.4 ± 2.6***</td>
<td>45.4 ± 1.9</td>
<td>33.4 ± 2.6***</td>
<td>48.2 ± 2.0</td>
<td>36.4 ± 2.7***</td>
</tr>
<tr>
<td>PM</td>
<td>52.0 ± 4.6</td>
<td>43.1 ± 6.7</td>
<td>58.6 ± 4.3</td>
<td>37.5 ± 5.6**</td>
<td>58.6 ± 4.5</td>
<td>52.8 ± 6.2</td>
</tr>
<tr>
<td>TM</td>
<td>35.3 ± 8.8</td>
<td>14.0 ± 7.4</td>
<td>43.1 ± 8.4</td>
<td>24.4 ± 7.9</td>
<td>53.7 ± 8.9</td>
<td>29.5 ± 8.4*</td>
</tr>
<tr>
<td>CYT</td>
<td>34.7 ± 5.4</td>
<td>29.6 ± 7.7</td>
<td>39.9 ± 5.2</td>
<td>24.9 ± 5.9</td>
<td>48.4 ± 5.6</td>
<td>41.0 ± 7.7</td>
</tr>
<tr>
<td>All</td>
<td>39.2 ± 1.4</td>
<td>28.8 ± 1.9***</td>
<td>43.5 ± 1.3</td>
<td>30.5 ± 1.7***</td>
<td>47.4 ± 1.4</td>
<td>35.3 ± 1.8***</td>
</tr>
</tbody>
</table>

**NOTE.**—See **NOTE** to table 5.

Adaptive Diversification of $\alpha$ Chains

In Table 7, mean values of $d_{\kappa}$ are presented for different gene regions, in comparisons among vertebrate integrin $\alpha$ chains with I-DOM, among vertebrate $\alpha$ chains without I-DOM, and between the two groups. In every case, MBR and TM are the most conserved regions. In most comparisons among different $\alpha$-chain classes, there was no evidence of a bias toward amino acid replacements changing residue properties in any region (data not shown). One exception involved the comparison between
Table 7
Mean $d_N \pm SE$ per 100 Nonsynonymous Sites, in Comparisons of Different Regions of Integrin $\alpha$-Chain Genes

<table>
<thead>
<tr>
<th>Region</th>
<th>$\alpha$ with I-DOM (N = 10)</th>
<th>$\alpha$ without I-DOM (N = 6)</th>
<th>$\alpha$ with I-DOM vs. $\alpha$ without I-DOM (N = 20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTR</td>
<td>72.7 ± 3.5***</td>
<td>85.0 ± 4.7</td>
<td>120.0 ± 5.5***</td>
</tr>
<tr>
<td>I-DOM</td>
<td>66.6 ± 3.7*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MBR</td>
<td>55.6 ± 3.5</td>
<td>75.9 ± 4.5</td>
<td>71.5 ± 3.4</td>
</tr>
<tr>
<td>CEN</td>
<td>93.2 ± 3.1***</td>
<td>101.0 ± 3.7***</td>
<td>129.0 ± 3.7***</td>
</tr>
<tr>
<td>TM</td>
<td>44.2 ± 6.8</td>
<td>68.0 ± 9.5</td>
<td>62.9 ± 7.3</td>
</tr>
<tr>
<td>All</td>
<td>77.9 ± 1.9</td>
<td>85.5 ± 2.6</td>
<td>113.6 ± 2.5</td>
</tr>
</tbody>
</table>

NOTE.—$N$ = Number of comparisons. Tests of significance of the difference between $d_N$ and that for MBR: * = $P < .05$; and *** = $P < .001$.

human $\alpha V$ and human $\alpha M$ and $\alpha X$. $\alpha M$ and $\alpha X$ are closely related (fig. 3), and both associate with $\beta 2$. $\alpha V$ differs in that it can associate with most $\beta$ chains except $\beta 2$. In CEN, charge is conserved between human and mouse $\alpha M$; but, between $\alpha V$ and $\alpha M/\alpha X$, nonsynonymous nucleotide substitutions causing charge change occur more frequently than expected by chance (table 8).

Discussion
Correspondences between $\alpha$- and $\beta$-Chain Phylogenies

The simplest model for coevolution of two multigene families whose products form heterodimers would involve coordinated duplication of genes within each family. As a result, the phylogeny of each gene family would exactly parallel that of the other; and any two genes whose products formed a heterodimer would occupy corresponding

Table 8
$p_{NC} \pm SE$ and $p_{NR} \pm SE$, with Respect to Amino Acid Residue Charge and Polarity, in Comparisons of Different Regions of Integrin $\alpha$-Chain Genes

<table>
<thead>
<tr>
<th>Characteristic and Region</th>
<th>HUMAN vs. MOUSE $\alpha M$</th>
<th>HUMAN $\alpha V$ vs. $\alpha M, \alpha X$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p_{NC}$</td>
<td>$p_{NR}$</td>
</tr>
<tr>
<td>Charge:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTR</td>
<td>14.8 ± 2.0</td>
<td>9.3 ± 2.1</td>
</tr>
<tr>
<td>MBR</td>
<td>11.2 ± 2.3</td>
<td>10.1 ± 2.7</td>
</tr>
<tr>
<td>CEN</td>
<td>17.0 ± 1.5</td>
<td>10.5 ± 1.6**</td>
</tr>
<tr>
<td>TM</td>
<td>8.6 ± 4.1</td>
<td>0.0 ± 0.0*</td>
</tr>
<tr>
<td>All</td>
<td>15.4 ± 1.1</td>
<td>10.1 ± 1.1**</td>
</tr>
<tr>
<td>Polarity:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTR</td>
<td>13.8 ± 1.9</td>
<td>10.6 ± 2.4</td>
</tr>
<tr>
<td>MBR</td>
<td>14.2 ± 2.5</td>
<td>5.0 ± 2.0**</td>
</tr>
<tr>
<td>CEN</td>
<td>16.1 ± 1.4</td>
<td>11.5 ± 1.7*</td>
</tr>
<tr>
<td>TM</td>
<td>8.2 ± 4.5</td>
<td>4.2 ± 4.1</td>
</tr>
<tr>
<td>All</td>
<td>15.1 ± 1.0</td>
<td>10.0 ± 1.2**</td>
</tr>
</tbody>
</table>

NOTE.—All data are multiplied by 100. Tests of significance of the difference between $p_{NC}$ and $p_{NR}$: * = $P < .05$; ** = $P < .01$; and *** = $P < .001$. 

positions in the phylogenies of their respective gene families. Such coordinated gene
duplication would probably only be possible if the two genes forming a heterodimer
were adjacent to each other on the chromosome and were duplicated in tandem. In
the class II major histocompatibility complex (MHC) of mammals, for example, the
α- and β-chain genes are adjacent; and, in the history of the class II MHC, tandem
duplications have played a role in generating new heterodimers (Hughes and Nei
1990). However, even in this case, a new heterodimer may arise by means other than
tandem duplication; for example, the DP heterodimer is composed of (1) a β chain
derived by duplication of a DQ β chain and (2) an α chain derived by duplication of
a DR α chain (Hughes and Nei 1990).

Nothing is so far known about the relative chromosomal arrangement of the
integrins genes of vertebrates, but the present analyses make clear that the α- and β-
chain gene families have not always evolved in a closely coordinated fashion. The
major division of vertebrate integrin β chains (between β1/β2 and β3/β5/β6) does
not correspond to the major division of vertebrate integrin α chains (between those
with I-DOM and those without). Furthermore, within each gene family there are
examples of highly generalist molecules (αV and β1) that form heterodimers, virtually
without regard for phylogeny. On the other hand, within each gene family there are
clusters of phylogenetically related molecules that form heterodimers with either the
same set of molecules from the other gene family or an overlapping set; for example,
αL, αM, and αX are closely related, and all associate with β2. Likewise, β3 and
β5 are closely related, and both form heterodimers that bind RGD-containing
ligands; both β3 and β5 associate with αV, and β3 also associates with the closely
related αIIb.

Rates of Evolution

In the absence of evidence regarding the time of divergence of different classes
of integrin genes, it is difficult to compare rates of evolution. However, in the case of
αM, β1, and β2, sequences are available for both mouse and human. \( d_N \) between
human and mouse β2 is about three times that between human and mouse β1 (table
4); and that between human and mouse αM (table 2) is more than four times that
between human and mouse β1. β1 thus appears to be extraordinarily conserved, in
comparison with other integrin α and β chains.

The degree of conservation between the two β1 chains of *Xenopus laevis* is still
more remarkable. Assuming both that the tetraploidization of *Xenopus* took place 40
Mya and that the human-mouse divergence took place 80–100 Mya, we expect the
ratio of \( d_N \) between human and mouse β1 to that between *X. laevis* β1a and β1b to
be 2:1–2.5:1. In fact, it is 4.25:1 (table 4). On the other hand, \( d_S \) between human
and mouse β1 is 61.7 ± 5.2, while that between *X. laevis* β1a and β1b is 31.2 ± 3.1.
Thus, in the case of synonymous substitutions, the ratio is very close to the expected
2:1. The fact that the rate of nonsynonymous evolution between the two *X. laevis*
β1 genes is much lower than the mammalian rate whereas the synonymous rate is not
shows that similarity of the two *X. laevis* proteins is maintained by strong purifying
selection rather than by concerted evolution. Both *X. laevis* β1 chains are apparently
expressed and associate with the same α chains (DeSimone and Hynes 1988). This
may impose particularly strong selection for them to maintain similar amino acid
sequences. The strength of conservation of charge between the two *X. laevis* β1 se-
quences is particularly remarkable; the ratio of \( p_{NC} \) with respect to charge to \( p_{NR} \) is
13:1, as opposed to 2:1 between mouse and human (table 5).
Interlocus Genetic Exchange

The integrin receptors have diversified to perform a wide variety of functions, and recombinational events have apparently played a role in this process. For example, the insertion of I-DOM into one lineage of α chains may have some functional significance. Integrins having I-DOM are functionally united, in that all appear to lack RGD-crosslinking function. Except for α4, vertebrate integrins without I-DOM do have RGD-crosslinking function. Presumably the transfer of I-DOM to α integrins and the transfer of a fibronectin-like repeat region to β4 both occurred by a process of exon shuffling, although the boundaries of I-DOM as determined by sequence alignment do not correspond closely to the current boundaries of exons in αX.

The donation of MBR from human αM to αX presumably also involved exon shuffling. Even in MBR, αX and αM may have been too divergent, especially at synonymous sites, for gene conversion (heteroduplex formation and repair) to occur (Walsh 1987). It is interesting that several recently reported cases of genetic exchange in multigene families resemble this one, in that a conserved domain has been exchanged between genes that otherwise are highly divergent. An example is the exchange of the highly conserved α3 domain among human CD1 antigen genes (Hughes 1991a). In these genes, adjacent exons encoding the α1 and α2 domains are subject to very little constraint and thus have diverged greatly among different members of this gene family (Hughes 1991a). Another example involves the α3 domain of the mouse class I MHC loci, where α1 and α2 have diverged as a result of positive selection (Hughes and Nei 1988).

Exchange of conserved domains may have adaptive significance, in that it serves to homogenize a region that has the same role in two related proteins. In the case of both CD1 and MHC, the α3 domain associates with β2-microglobulin. Since β2-microglobulin is the product of a single locus but forms heterodimers with several distantly related molecules, homogeneity in the α3 domain may be selectively favored. Both αM and αX integrins interact with β2 integrin, and MBR may play a role in this interaction.

Positive Selection on Charge Profiles

Analysis of conservative and radical nonsynonymous nucleotide substitutions in different regions of integrin β-chain genes (tables 5 and 6) showed a bias toward charge change in CEN of functionally divergent molecules. Such a pattern is likely to be the result of positive selection favoring divergence of charge profile in this region.

These results pinpoint CEN as a likely focus of adaptive diversification among classes of integrin β chains and suggest that charge profile in this region may be adapted to the specific function of each class of β chains. Given the importance of charged residues in association between subunits of dimeric proteins (Miller et al. 1989), it is possible that charge profile in CEN plays a role in the association between each β chain and the array of α chains with which it interacts. Some results are consistent with this explanation. For example, the two X. laevis β1 chains are believed to bind the same α chains (DeSimone and Hynes 1988) and after 40 Myr have no charge differences in CEN (table 6).

Likewise, it is also possible that the charge-change bias between human αV and αM/αX represents coadaptation of these α chains to the different sets of β chains.
with which they associate. However, until the detailed structure of the integrin receptor is known, it remains uncertain which domains are involved in interactions between α and β chains. It is possible that in CEN the adaptive charge-profile divergence among different classes of integrin β chains relates to some factor other than association with a given set of α chains; for example, it may be an adaptation to the set of ligands bound by each integrin receptor.

Since most nonsynonymous mutations are deleterious, they are eliminated by purifying selection. As a consequence, $d_S$ generally exceeds $d_N$ (Li et al. 1985). When balancing selection favors diversity at the amino acid level, a reversal of this pattern can be seen; thus a comparison of $d_S$ and $d_N$ can provide a test for positive selection in such cases (Hughes and Nei 1988). Directional positive selection leading to functional divergence among a set of related proteins can produce a similar acceleration of $d_N$, as Hill and Hastie (1987) found in the case of serine protease inhibitors. In the latter case, however, some purifying selection apparently has subsequently occurred in the same region as directional selection; as a result, in some comparisons $d_S$ exceeded $d_N$, even in the region where directional selection was supposed to have taken place (Hill and Hastie 1987).

In most cases where, as the result of directional positive selection, members of a multigene family have diverged adaptively in a certain region, this has happened either in the distant past or gradually over a long period of time. Consequently, the effects of purifying selection generally mask those of positive selection, making it impossible to distinguish between the latter and a lack of constraint on the amino acid sequence. Under such circumstances, the present method can identify regions where directional positive selection has acted. As long as $d_N$ is not very high, evidence of directional change in amino acid residue characteristics can be detected by this method, even after several hundred million years.

Acknowledgment

This research was supported by a grant from the National Institutes of Health.

LITERATURE CITED


PYTELA, R. 1988. Amino acid sequence of the murine Mac-1 α chain reveals homology with the integrin family and an additional domain related to von Willebrand factor. EMBO J. 7:1371–1378.


JAN KLEIN, reviewing editor

Received February 28, 1991; revision received September 23, 1991

Accepted September 23, 1991
Phylogenetic Position of Phylum Nemertini, Inferred from 18S rRNA Sequences: Molecular Data as a Test of Morphological Character Homology

J. McClintock Turbeville,* Katharine G. Field,† and Rudolf A. Raff*

*Department of Biology and Institute for Molecular and Cellular Biology, Indiana University; and †Department of Microbiology, Oregon State University

Partial 18S rRNA sequence of the nemertine Cerebratulus lacteus was obtained and compared with those of coelomate metazoans and acoelomate platyhelminths to test whether nemertines share a most recent common ancestor with the platyhelminths, as traditionally has been implied, or whether nemertines lie within a protostome coelomate clade, as suggested by more recent morphological analyses. Maximum-parsimony analysis supports the inclusion of the nemertine within a protostome-coelomate clade that falls within a more inclusive coelomate clade. Bootstrap analysis indicates strong support for a monophyletic Coelomata composed of a deuterostome and protostome-coelomate clade. Support for a monophyletic protostome Coelomata is weak. Inference by distance analysis is consistent with that of maximum parsimony. Analysis of down-weighted paired sites by maximum parsimony reveals variation in topology only within the protostome-coelomate clade. The relationships among the protostome coelomates cannot be reliably inferred from the partial sequences, suggesting that coelomate protostomes diversified rapidly. Results with evolutionary parsimony are consistent with the inclusion of the nemertine in a coelomate clade. The molecular inference corroborates recent morphological character analyses that reveal no synapomorphies of nemertines and flatworms but instead suggest that the circulatory system and rynchocoel of nemertines are homologous to coelomic cavities of protostome coelomates, thus supporting the corresponding hypothesis that nemertines belong within a protostome-coelomate clade. The sequence data provide an independent test of morphological character homology.

Introduction

Morphological and embryological data have contributed substantially to the elucidation of the phylogenetic relationships of animal phyla. However, the number of different phylogenies proposed for the Metazoa reflects the inability of zoologists to arrive at a consensus view of metazoan interphyletic relationships based on these characters (e.g., see Hyman 1940, p. 38; Ax 1984, pp. 285–286, 1989; Brusca and Brusca 1990, pp. 879–889). There are several reasons for the uncertainty regarding metazoan relationships. First, most phyla share few informative homologous anatomical or embryological features. Second, it is often exceedingly difficult to distinguish homologous similarity (homology) from convergent or parallel similarity (analogy),...
and this is a prerequisite for inferring phylogenetic relationships. Third, the fossil record has not provided informative characters for linking most phyla, and many small and soft-bodied phyla lack a fossil record. Sequence data obtained from ribosomal RNAs offer an important new source of informative characters for inferring high-level phylogenetic relationships for many taxa and provide an independent test of hypotheses based on morphological characters, especially for metazoans (Woese 1987; Field et al. 1988; Abele et al. 1989).

We report here our investigation of the 18s rRNA of Cerebratulus lacteus from the phylum Nemertini or Rhynchocoela, a group of worms whose relationships have been controversial. Nemertines, or ribbon worms, are primarily epifaunal or infaunal inhabitants of the marine benthos, but marine pelagic, freshwater, and terrestrial forms also occur. The phylum comprises ~900 species (Gibson 1982). Nemertines are typically a few millimeters to several centimeters in length, but many species are much longer; there are reports in the classical literature of a species that attains a length of 30 m (Bürger 1897–1907). The major diagnostic feature of the phylum is an eversible proboscis enclosed when uneverted in a fluid-filled, cell-lined cavity, the rhynchocoel. In addition to this secondary body cavity, nemertines possess a continuous system of cell-lined channels historically referred to as a blood vascular system (Bürger 1897–1907; Hyman 1951, p. 486).

Nemertines traditionally have been considered acoelomate animals most closely related to the platyhelmintths, or flatworms, on the basis of morphological and limited immunological similarity (among others, see Bürger 1897–1907; Schepotieff 1912; Hyman 1951, p. 528; fig. 1A). This hypothesis has been perpetuated by modern biology and invertebrate zoology textbooks and by more detailed accounts of the phylum (e.g., see Gibson 1972, pp. 187–191; Lutz 1986, p. 201; Curtis and Barnes 1989, p. 538; Willmer 1990). However, recent structural and ultrastructural analyses corroborate an alternative, minority hypothesis put forward by earlier zoologists (e.g., see Nusbaum and Oxner 1913; Friedrich 1935). The data suggest that the nemertine circulatory system and rhynchocoel are homologous to coelomic cavities of protostome coelomates (also referred to as schizocoelous or spiralian coelomates) and thus support the hypothesis that nemertines are most closely related to protostome coelomates such as annelids and mollusks and that they thus belong within a protostome-coelomate clade (fig. 1B; Turbeville 1986b; Turbeville 1991). The interpretation of nemertine body cavities as coelom homologues—and this alternative hypothesis of nemertine relationships—have not been accepted by all invertebrate systematists, further reflecting the difficulty of inferring the phylogenetic relationships of nemertines by use of morphological characters (e.g., see Bartolomaeus 1988). Additional informative characters are required to clarify the phylogenetic relationships of nemertines. The objective of the present investigation is to test hypotheses of nemertine relationships by using 18S rRNA sequence data with maximum-parsimony, distance-matrix, and evolutionary-parsimony methods of analysis.

Material and Methods

Organisms Analyzed

Eleven sequences representing six metazoan phyla were compared in the present study. These were Dugesia tigrina (Platyhelminthes), Bothromesostoma personatum (Platyhelminthes), Fasciola hepatica (Platyhelminthes), Chaetopterus variopedatus (Annelida), Lumbricus species (Annelida), Cerebratulus lacteus (Nemertini), Cryptochiton stelleri (Mollusca), Golfoingia gouldii (Sipunculida), Asterias forbesii (Echi-
Fig. 1.—Conflicting hypotheses of nemertine relationships based on morphological characters. A, Orthodox or traditional hypothesis, indicating monophyly of Nemertini + Platyhelminthes. No unequivocal synapomorphies support this hypothesis (see text). B, Alternative hypothesis, indicating that nemertines belong within schizocoelous (see table 1) or protostome coelomate clade that is part of more inclusive coelomate clade. For an alternative hypothesis, see Brusca and Brusca (1990, pp. 331 and 882).
noderma), and Branchiostoma californiense (Chordata). With the exception of the Cerebratulus sequence reported here and Bothromesostoma and Fasciola sequences (M. Riutort and K. G. Field, personal communication), all sequences were collected by Field et al. (1988). The nemertine and the new flatworm sequences have been submitted to EMBL.

RNA Extraction and Sequencing

Specimens of Cerebratulus lacteus Leidy 1851 were obtained from the Marine Biological Laboratory (Woods Hole, Mass.). Total RNA was extracted from oocytes by following a protocol adapted from Paterson and Roberts (1981) and using 8 M guanidine hydrochloride (Turbeville et al. 1991). Direct sequencing of 18S rRNA was accomplished by employing the method of Lane et al. (1985, 1988), which is a modification of the Sanger et al. (1977) dideoxynucleotide chain-terminating technique using reverse transcriptase. Six oligonucleotide primers complementary to specific conserved regions of the molecule were used in separate reactions. Sequences of the primers utilized were as follows (numbers correspond to positions on the human sequence): 445–429, 5'-TCAGGCTCCCTCTCCGG-3'; 632–615, 5'-GWAT-TACCGCGCGKGTGCTG-3'; 880–865, 5'-CCGAGGCTCATTCCA-3'; 1009–993, 5'-TTGGCAAATGCTTTCGC-3'; 1201–1187, 5'-ATCCTTTRAGTTTC-3'; and 1708–1692, 5'-GACCGGCGGTGTGTRCA-3'.

Sequence Alignment

Sequence data were aligned by hand, beginning at universally conserved regions. Alignment gaps were inserted to account for putative length differences between the sequences (fig. 2). Alignments were refined using the secondary-structure models inferred for the bird spider Eurypelma californica (Hendriks et al. 1988). Positions aligned were limited to regions for which sequence data are available for all taxa. Of the 1,140 aligned positions, 946 were included in the analyses. Positions exhibiting high variability or length variation could not be reliably aligned and were excluded from the analyses. Positions analysed are indicated in figure 2.

Data Analysis

Data were analyzed using the maximum-parsimony method of the PAUP 3.0d package written by D. L. Swofford (Illinois Natural History Survey), the least-squares distance-matrix program of Olsen (1988a, 1988b), and a phylogenetic-invariants method (Lake 1987). For parsimony analysis, characters were entered unordered, and gaps were treated as missing data. The BRANCH AND BOUND option of PAUP, which guarantees the most parsimonious solutions for a given data set, was used in all analyses except bootstrapping (see below). The analysis was performed with all nucleotides and was repeated with paired positions downweighted by one-half, following the suggestion of Stahl et al. (1984) and Wheeler and Honeycutt (1988). The latter authors postulate that, the paired regions might be less reliable than unpaired regions, for inferring phylogeny, as is the case for 5S rRNA. Downweighting was accomplished in practice by assigning the unpaired positions twice the weight of the paired sites. Paired and unpaired positions were inferred by juxtaposing primary structures with the secondary-structure model proposed by Hendriks et al. (1988) for the spider Eurypelma. The cnidarian Hydra was chosen for the outgroup, on the basis of morphological and molecular analyses that suggest that the Cnidaria is the sister taxon of the bilateral metazoans (Ax 1989; authors’ unpublished data). A bootstrap analysis for
placing confidence intervals on inferred phylogenies was also utilized (Felsenstein 1985), and 100 replications were run using the total data set. For bootstrapping, the HEURISTIC search was employed by using the TBR branch-swapping algorithm in combination with CLOSEST stepwise addition. Twenty trees were held at each step.

Data were analyzed by the distance-matrix-analysis program of Olsen (1988a, 1988b), which is based on the method developed by Fitch and Margoliash (1967); the Jukes and Cantor (1969) correction to estimated distances was employed. In the present study, trees were rooted by the most distantly related metazoan, *Hydra*. The bootstrap algorithm was also applied to the distance analysis. One hundred replications were run. The analysis was also run with downweighted paired positions, as described above.

The method of evolutionary parsimony (Lake 1987) was applied to all positions. The version compiled for the PAUP 3.0 package was used. For evolutionary-parsimony analyses, sequences were divided into four groups, and the three possible topologies for all quartets composed of a single sequence from each group were evaluated, and the results were combined. The $\chi^2$ values for combined trees were calculated by following the method presented by Lake (1987, Appendix). When necessary, correlation values were corrected as explained by Turbeville et al. (1991), prior to calculating the $\chi^2$ values.

Results

Direct sequencing of the 18S rRNA molecules of *Cerebratulus lacteus* by using six primers yielded 1,208 nucleotides. Unambiguous nucleotide assignments could not be made at 53 sites, or 4.4% of the total. Ambiguities result from common artifacts of reverse-transcriptase sequencing (see Lane et al. 1985). Of the nucleotides obtained, 1,140 are shown in the alignment (fig. 2). A unique insertion (autapomorphy) 20 nucleotides in length is present between positions 200 and 202 (fig. 2).

Maximum-Parsimony Analysis

For the taxa analyzed, 253 of the reliably aligned variable sites (fig. 2) are informative for inferring phylogeny by maximum-parsimony analysis. Analysis of all positions found a single minimum-length tree of 874 steps (fig. 3A). Maximum-parsimony analysis indicates that the nemertine *Cerebratulus lacteus* falls within a protostome or spiralian coelomate clade as the sister group of *Golgingia* rather than being the sister group of the acoelomate Platyhelminthes (fig. 3A). The *Cerebratulus*-plus-*Golgingia* clade is the sister taxon of a *Cryptochiton*-plus-*Chaetopterus* clade. The oligochaete annelid *Lumbricus* is the sister group of these other protostome-coelomate taxa. Thus, monophyly of the Annelida is not supported by the partial sequence data (see Discussion). The reliability of the phylogeny was estimated by bootstrap analysis of the entire data set. Ninety-six percent of the trees from 100 bootstrap replications supported monophyly of the coelomates inclusive of the nemertine (fig. 3A). Monophyly of the spiralian coelomates is supported to a lesser extent, and relationships among the taxa of this clade are weakly supported (fig. 3A). To assess further the reliability of the maximum-parsimony inference, we considered the 124 trees saved within 1% of the length of shortest tree (874–883 nucleotidic substitutions). A 50%-majority-rule consensus tree is shown in figure 3B. Variation in topology was observed within the protostome-coelomate and platyhelminth clades. When paired positions are downweighted and the total data set is analyzed, a single tree is found that has a topology that varies only within the protostome-coelomate clade (not shown). The
FIG. 2.—Partial 18S rRNA sequences aligned to corresponding regions of *Homo sapiens* (Chordata; Torczynski et al. 1983) sequence for reference. Only regions for which sequence data are available for all taxa are included. V = Position used in analyses. Ho = *Homo*; Br = *Branchiostoma*; As = *Asterias*; Go = *Golgi*; Ce = *Cerebratulus*; Ch = *Chaetopterus*; Cr = *Cryptochiton*; Lu = *Lumbricus*; Bo = *Bothromesostoma*; Du = *Dugesia*; Fa = *Fasciola*; and Hy = *Hydra*. Accession numbers for unpublished sequences are as follows: Ce, M81167; Bo, M58347; and Fa, X56041.
FIG. 3.—Trees inferred by maximum-parsimony analysis. A, Shortest tree (874 nucleotide substitutions). The overall consistency index is 0.708. The consistency index when uninformative positions are excluded is 0.605. Numbers at nodes indicate the frequency with which the clade descending from that node was found by bootstrapping. Branch lengths are proportional to the number of nucleotide substitutions. The scale bar indicates the approximate number of nucleotide substitutions. B, Fifty-percent-majority rule consensus tree of 124 trees saved within 1% length of most parsimonious tree.
nemertines become the sister group of the rest of the protostome coelomates. Within the subordinate clade, the polychaete *Chaetopterus* is placed between the chiton and a *Lumbricus*-plus-*Golfingia* clade. If maximum-parsimony analyses were run with the nucleotides excluded from the analyses, two equally parsimonious trees were found with variation only within the protostome-coelomate clade (authors' unpublished results).

**Distance Analysis**

Distance-matrix results are consistent with those of the maximum-parsimony analyses, indicating that nemertines lie within a protostome-coelomate group and that the nemertine RNA molecule has accumulated a somewhat greater number of substitutions than have those of the other protostome coelomates included in the analysis (fig. 4). The topology of the tree inferred from all positions agrees with the parsimony analysis, except among the protostome coelomates. The sipunculid is the sister group of the rest of the protostome-coelomate taxa, and the nemertine is placed between the oligochaete *Lumbricus* and a *Cryptochiton*-plus-*Chaetopterus* group. The distance tree inferred from downweighting paired positions is identical in topology to the tree inferred from unweighted data. Bootstrap analysis of the total data set indicates support both for the coelomate lineage in 100% of the outcomes and for a protostome-coelomate group in 85% (fig. 4). Support for relationships among the protostome coelomates is

![Distance tree](image)
somewhat weaker (fig. 4). No variation in tree topology was observed (authors' unpublished results) when unreliably aligned positions were included in the analysis.

Evolutionary Parsimony

For evolutionary parsimony, the total number of informative transversion positions (parsimony counts minus background counts) per quartet was between zero and nine. Figure 5 illustrates inferences of nemertine relationships by evolutionary parsimony. The results are a subset of all possible combinations. The first set of analyses tests whether the nemertine is most closely related to protostome coelomates, deuterostomes, or flatworms. The favored tree \( (P \sim 0.3) \) in the first set of analyses (fig. 5A) indicates monophyly of nemertine + protostome coelomates, although support is not significant (fig. 5A). The results are also dependent on which taxa are included in the four groups (see Turbeville et al. 1991). The \( P \) values of the two alternative topologies, one linking the nemertines and the flatworms and the other linking the nemertine and deuterostomes, are \( P = 1 \) and \( P \sim 0.5 \), respectively. The second set of analyses was designed to test monophyly of the coelomates and the nemertine (fig. 5B). The favored tree links the coelomate taxa and the nemertine, but support for this topology is also not significant \( (P \sim 0.07) \).

Discussion

The results support the hypothesis that nemertines are coelomate animals that belong within a protostome-coelomate clade. Measures of reliability, including bootstrap analyses and consideration of all trees within 1% of the length of the most parsimonious solution, firmly support coelomate monophyly and, to a lesser extent, protostome-coelomate monophyly. Results with evolutionary parsimony are less robust, perhaps owing to the small number of informative transversion positions (between

![Figure 5](image_url)

**Fig. 5.—** Summary of analyses with evolutionary parsimony. A, Test of nemertine+protostome-coelomate monophyly. The favored topology resulting from combining results of 24 quartets \((1 \times 2 \times 3 \times 4)\) is illustrated. One taxon from each group was compared in turn. The favored tree links the nemertine (Cerebratulus) to the protostome coelomates Lumbicus, Chaetopterus, Golfingia, and Cryptochiton \((P \sim 0.3)\) rather than to the platyhelminths Dugesia, Bothromesostoma, and Fasciola \((P = 1)\) or to the deuterostomes Asterias and Branchiostoma \((P \sim 0.5)\). B, Test of nemertine+coelomate monophyly. Eighteen quartets \((1 \times 1 \times 3 \times 6)\) were evaluated and combined. The favored topology is shown and links the nemertine and coelomates. \( P \) values for the tree linking the nemertine and flatworms and for the tree linking the nemertine and cnidarian are, respectively, \( P \sim 0.8 \) and \( P \sim 0.4 \). The \( \chi^2 \) test for correlated data presented by Lake (1987) was used for both trees.
zero and nine) in the data set. All analyses support monophyly of coelomates and are consistent, in part, with the studies of Field et al. (1988), Ghiselin (1988), and Lake (1990). The coelomates or "Coelomata" comprise the deuterostome and protostome-coelomate clades (figs. 1B-4). Monophyly of Platyhelminthes is also firmly supported by the partial sequence data. The present study and those by Field et al. (1988) and Ghiselin (1988) suggest that Platyhelminthes is the most primitive bilateral metazoan taxon and thus constitutes the sister group of the remaining Bilateria or Eubilateria. These preliminary molecular analyses support, in part, one of two equally plausible hypotheses based on a cladistic analysis of morphological features (Ax 1984, pp. 258-286, 1989).

Downweighting paired sites of 18S rRNA altered topology within the protostome-coelomate clade only when they were analyzed by maximum parsimony, suggesting that the paired sites contain some noise. However, the utility of this procedure is uncertain (see Smith 1989; Hedges et al. 1990).

Measures of reliability indicate that relationships among the protostome-coelomate taxa cannot be confidently inferred by the partial sequence data; even the Annelida is not inferred as monophyletic. These results may be attributable, at least in part, to the fact that the regions of the molecule sequenced and compared are the most highly conserved. For example, the polychaete annelid (Chaetopterus) sequence and the nemertine sequence are ~90% identical. It is reasonable to assume that relatively few substitutions have occurred in the molecule after the protostome-coelomate clade became distinct and before the taxa of the protostome-coelomate clade diverged, and this situation limits the number of potentially informative sites available for inferring their precise relationships. This suggests rapid diversification of these taxa after the initial radiation of the coelomates. When one is inferring relationships at great evolutionary distances, it is important to limit analyses to well-conserved (i.e., slowly evolving) molecules or regions of a molecule, in order to minimize the amount of homoplasy in the data set. Sequence data from the rest of the 18S rRNA molecule and from other more rapidly evolving genes should offer a greater number of informative positions for inferring more recent divergences.

There exist two dominant hypotheses of nemertine relationships based on morphological characters. The first, or orthodox, view states that nemertines are most closely related to platyhelminths (flatworms), the implication being that nemertines and flatworms shared a most recent common ancestor (fig. 1A). The second hypothesis holds that nemertines are coelomate worms that belong within a protostome-coelomate clade (fig. 1B). Body organization, specifically the organization of the space between the body wall and gut, is of central importance for understanding the conflicting hypotheses of nemertine relationships and must first be considered.

Historically, nemertines have been regarded as acoelomate in body organization, a condition considered homologous to that of the Platyhelminthes. This shared similarity has been considered support for the hypothesis that nemertines and platyhelminths shared a most recent common ancestor. However, it was also known that nemertines possessed both a cell-lined cavity enclosing an eversible proboscis, the rhynchocoel, and an independent system of cell-lined cavities forming a continuous loop, or circulatory system. Proponents of the orthodox hypothesis (e.g., see Bürger 1897-1907; Hyman 1951, pp. 486-490; Gibson 1972, pp. 71-75) implicitly considered the rhynchocoel to be a unique coelom analogue and apparently viewed the circulatory vessels as blood-vessel homologues. However, a minority of investigators suggested that these spaces are coelom homologues and that nemertines thus are actually coelo-
mate organisms (e.g., see Nusbaum and Oxner 1913; Nawitzki 1931; Friedrich 1935). The latter hypothesis has, for the most part, been ignored in the English zoological literature, perhaps because of the language barrier (the articles are in German) and because limited and sometimes conflicting data on vessel ontogeny did not clearly corroborate anatomical and histological evidence of vessel and coelom homology (for extensive review, see Turbeville 1986a, 1986b). The hypothesis that nemertine vessels are coelom homologues has been tested recently by utilizing transmission-electron microscopy, which enhances the evaluation of characters and thereby allows for more critical tests of hypotheses of homology by following established criteria of homology recognition (table 1; Turbeville 1986a, 1986b, 1991).

Ultrastructural analyses of adult vessel morphology and an analysis of vessel ontogeny provided evidence supporting the alternative hypothesis that the circulatory-system vessels are coelom homologues, thus suggesting that nemertines belong within a protostome-coelomate clade. In anatomical position, histology, cytology, and mode of formation, the vessels correspond to coelomic cavities of coelomate protostomes, whereas the position, composition, and formation of invertebrate blood vessels are considerably different (table 1; Turbeville 1986b, 1991). Cell-lined vertebrate vessels also are structurally and developmentally unlike nemertine vessels (for discussion, see Turbeville 1986b). Data on the rhynchocoel also are consistent with its interpretation as a modified coelom homologue (Turbeville 1991). The molecular data corroborate the morphological analysis of nemertine vessels and rhynchocoel, supporting the alternative hypothesis. Some invertebrate systematists (e.g., see Ax 1984, p. 271; Bartolomaeus 1988) interpret the correspondence of structure and ontogeny of nemertine vessels and coelomic cavities as convergent or parallel similarity and consider the vessel system in nemertines to be a unique feature (autapomorphy) of the phylum, thus rejecting the hypothesis of a close relationship between nemertines and coelomates. However, the molecular data do not support this interpretation.

In addition to a shared acoelomate condition that is unsupported by the data (see above), the orthodox hypothesis—i.e., that nemertines are most closely related

---

**Table 1**

<table>
<thead>
<tr>
<th></th>
<th>Nemertine Vessels</th>
<th>Coelomic Cavities</th>
<th>Invertebrate Blood Vessels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anatomical position</td>
<td>Lateral</td>
<td>Lateral</td>
<td>Dorsal and ventral</td>
</tr>
<tr>
<td>Lining</td>
<td>Continuous mesothelium</td>
<td>Continuous mesothelium</td>
<td>Lined by extracellular matrix or discontinuous</td>
</tr>
<tr>
<td></td>
<td>with intercellular junctions, cilia,</td>
<td>with intercellular junctions, cilia, and myofilaments</td>
<td></td>
</tr>
<tr>
<td></td>
<td>myofilaments*</td>
<td></td>
<td>cell layer lacking junctions cilia and myofilaments</td>
</tr>
<tr>
<td>Tissues separating</td>
<td>Mesoderm</td>
<td>Mesoderm</td>
<td>Extracellular matrices</td>
</tr>
<tr>
<td>during ontogeny</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note.**—Homology analysis follows established criteria for testing hypotheses of morphological character homology (see Remane 1956, pp. 30–60; Riedl 1978, pp. 33–36; Wiley 1981, pp. 130–138; Rieger and Tyler 1985).


*One unconfirmed report of myofilaments (see Turbeville 1986b).

This mode of coelom formation is referred to as schizocoely.
to flatworms—has been based on their common possession of several other presumably homologous characters, including the multiciliated epidermis, special secretory bodies termed “rhabdites,” ocelli, frontal organs, and nephridia (= excretory organs). Recent detailed analyses of these characters do not support their interpretation as shared derived homologues (synapomorphies; Turbeville 1991). Some of these characters (e.g., epidermis and nephridia) are symplesiomorphies, and the homology of others (e.g., frontal organs and rhabdites) is unsupported. Therefore, these characters do not support the hypothesis of a most recent common ancestry of platyhelminths and nemertines (see Bartolomaeus 1985, 1988; Turbeville and Ruppert 1985; Turbeville 1991). The molecular data are in full accord with this conclusion (figs. 3 and 4).

The microcomplement-fixation analysis of Schepotieff (1912) revealed immunological similarity between a single nemertine and a polyclad flatworm, rather than between the nemertine and a polychaete annelid, suggesting a closer relationship between nemertines and flatworms. However, both morphological and sequence analyses suggest that this shared similarity, if homologous, should be interpreted as a symplesiomorphy.

The molecular data support the hypothesis that nemertines are coelomate animals that belong within a protostome-coelomate clade (figs. 1B, 3 and 4). Thus, these data also affirm the interpretation of nemertine body cavities as coelom homologues rather than as independently derived coelom analogues. The sequence data have provided an independent but indirect test of morphological character homology. The results of both morphological and sequence analyses necessitate rejection of the hypothesis that nemertines represent the sister group of the flatworms. Furthermore, the molecular data refute the recently revived hypothesis that nemertines share a most recent common ancestry with vertebrates or chordates (Jensen 1988), a hypothesis also untenable on morphological grounds. The sequence data in the present paper thus provide informative characters allowing for the resolution of controversies resulting from differing interpretations of the same morphological characters. This study illustrates the importance of considering both molecular and morphological data when one is evaluating high-level evolutionary relationships of metazoans.

Acknowledgments

We thank Drs. B. Parr, N. Pace, T. Schmidt, and S. Turner for technical advice and L. Jaffee for assistance obtaining oocytes. We are grateful to Dr. M. Riutort for use of unpublished sequence data and to Dr. C. Marshall for a critical review of the manuscript. This study was supported by an Alfred P. Sloan Foundation Postdoctoral Fellowship to J.M.T. and by NSF Grant BSR-8818044 to R.A.R.

LITERATURE CITED


No direct contact between the excretory system and the circulatory system in Prostomatella arenicolu Friedrich (Nemertini). Hydrobiologia 156:175–181.


WALTER M. FITCH, reviewing editor

Received April 10, 1991; revision received August 28, 1991

Accepted September 25, 1991
Molecular Relationships between Alcohol Dehydrogenase Null-Activity Alleles from Natural Populations of *Drosophila melanogaster*

John B. Gibson, * Ann Verona Wilks, * and Alex Agrotis†

*Molecular and Population Genetics Group, Australian National University; and†Department of Genetics and Developmental Biology, Monash University

Alcohol dehydrogenase null-activity alleles extracted from a number of natural populations of *Drosophila melanogaster* in Tasmania were shown to be molecularly similar by probing, with an oligonucleotide specific to an inserted region in intron 2 of the gene, genomic DNA amplified by the polymerase chain reaction. This insertion had previously been shown to be the cause of the loss of activity in one of the null alleles whose DNA sequence was known. Three Adh null alleles from mainland populations did not contain the insertion. Two of these null alleles, extracted from the Coffs Harbour population in different years, were cloned, and their DNA sequences showed that they were identical and that both had a 438-bp deletion which removed most of exon 2. The third null allele, identified in a sample of flies from Chateau Tahbilk, was shown by 4-bp restriction-endonuclease mapping to contain a 320-bp insertion in intron 1, although this may not be the cause of the loss of activity. The data show that at least three different Adh null alleles have been found in Australian populations and that at least two have been maintained as heterozygotes over a period of years.

Introduction

The very extensive surveys of genetic variation in natural populations, carried out over the past 30 years, have occasionally uncovered null-activity alleles at structural gene loci. Freeth (1986), reviewing these occurrences, noted that null alleles had been reported for at least one structural gene in studies of 34 different plant and animal species. The prevalence of null alleles is particularly well documented in human populations, in which more than 100 metabolic diseases due to null- or low-activity variants have been described (Raivio and Seegmiller 1972; Mohrenweiser 1981).

The frequencies of null-activity alleles at structural gene loci in natural populations mainly fall within the range that can be explained by the balance between mutation and selection. In *Drosophila melanogaster*, mutation rates to null alleles have been estimated to be $1.03 \times 10^{-5}$ over five loci (Mukai and Cockerham 1977) and $3.86 \times 10^{-6}$ over seven loci (Voelker et al. 1980), while in newborn infants the frequency of enzyme-deficiency variants has been estimated to be 0.0024 (Mohrenweiser 1981). Specific studies aimed at assessing the frequency of null alleles in natural populations have observed values of 0.0031 over 29 loci in *Pinus ponderosa*. 0.0028 over 27 loci

1. Key words: alcohol dehydrogenase gene, *Drosophila melanogaster*, null alleles, DNA sequence.

Address for correspondence and reprints: John B. Gibson, Molecular and Population Genetics Group, Research School of Biological Sciences, The Australian National University, P.O. Box 475, Canberra, A.C.T. 2601, Australia.

© 1992 by The University of Chicago. All rights reserved.
0737-4038/92/0902-0005$02.00
in *P. resinosa* (Allendorf et al. 1982), and 0.0024 over 20 autosomal loci in *D. melanogaster* (Langley et al. 1981).

Exceptions to these low values have been reported, but they generally refer to null alleles at duplicated loci (Allendorf et al. 1984) where one copy of the gene retains a normal level of activity. An unusually high frequency of null-activity alleles at a single autosomal structural gene locus has recently been observed by Freeth and Gibson (1985) in Tasmanian (Australia) populations of *D. melanogaster*. They reported that the frequency of null-activity alleles at the alcohol dehydrogenase (ADH) locus (*Adh*) was 3.9% in one natural population and that in other populations, ≤300 km away on the same island, frequencies were >1%.

Comparisons of the biochemical properties of some of the *Adh* null alleles isolated from Tasmanian populations showed that they were similar to each other in that none of them produced ADH-CRM (cross-reacting material), yet they were transcribed, giving a low level of an mRNA ~200 bases longer than normal (Freeth et al. 1986, 1988). This apparent homogeneity might be an indication that the null alleles are molecularly the same, although the similarity in properties could be caused by different structural changes in the gene. If null alleles extracted from different Tasmanian populations contain the same molecular lesion, it would imply either that they are identical by descent and have spread from the population in which they first arose or, less probably, that the same mutation is recurrent. The DNA sequences of two of the Tasmanian null alleles, one from Huonville, *Adh*\(^{nAH52}\), and the other from Cygnet, *Adh*\(^{nAC14}\), revealed that both alleles contained the same alteration to intron 2, with eight extra bases inserted as two groups of four (Gibson and Wilks 1989; Freeth et al. 1990). In addition, *Adh*\(^{nAC14}\) contained a stop codon, TAG, in exon 2; however, the results of S1 nuclease protection experiments suggested that the structural change in intron 2 was responsible in both null alleles for the altered transcription patterns (Freeth et al. 1990).

Following the detection of null alleles in the Tasmanian populations, we have found *Adh* null alleles in surveys of a number of other Australian populations, albeit at lower frequencies than those found in Tasmania. Also, in two separate laboratories, null alleles have turned up among relatively small samples of *Adh* alleles isolated from natural populations. To test the similarities of these null alleles we have examined the molecular structures of null alleles isolated from four Tasmanian populations and compared them with *Adh* null alleles isolated from three mainland populations. The results show that molecularly different *Adh* null alleles are present in Australian populations of *D. melanogaster* and that they have arisen from both *Adh*\(^{E}\) and *Adh*\(^{S}\).

Material and Methods

The naturally occurring *Adh* null alleles we have investigated are listed in table 1. They are identified by the initial letter(s) of the population from which they were extracted and, if appropriate, by the single female line in which the allele was detected.

The Tasmanian null alleles investigated were taken at random from ~30 that had been extracted from four natural populations on different occasions between 1983 and 1986. Because of the ephemeral nature of *Drosophila melanogaster* breeding sites, samples from the same locality might be collected some distance apart (although usually not >4 km) in different years. For example, the samples from Coffs Harbour that were collected in 1984 and 1986 were from different banana plantations ~2 km apart, whereas at Cygnet (Tasmania) the collections were made in different areas of an apple-processing factory. All of the null alleles lacked ADH activity both in adults
Table 1
Adh Null Alleles Isolated from Australian Natural Populations of Drosophila melanogaster

<table>
<thead>
<tr>
<th>Adh Null Allele</th>
<th>Year Isolated</th>
<th>Habitat (location)</th>
</tr>
</thead>
<tbody>
<tr>
<td>nAdC.14</td>
<td>1983</td>
<td>Apple-processing waste (Cygnet, Tasmania)</td>
</tr>
<tr>
<td>nAdC.22</td>
<td>1983</td>
<td>Apple-processing waste (Cygnet, Tasmania)</td>
</tr>
<tr>
<td>nAdC.45</td>
<td>1984</td>
<td>Apple-processing waste (Cygnet, Tasmania)</td>
</tr>
<tr>
<td>nAdP.91</td>
<td>1983</td>
<td>Grape must (Pipers Brook, Tasmania)</td>
</tr>
<tr>
<td>nAdT.33</td>
<td>1983</td>
<td>Grape must (Tamar, Tasmania)</td>
</tr>
<tr>
<td>nAdT.54.0</td>
<td>1984</td>
<td>Grape must (Tamar, Tasmania)</td>
</tr>
<tr>
<td>nAdT.52</td>
<td>1985</td>
<td>Waste apples (Huonville, Tasmania)</td>
</tr>
<tr>
<td>nAdT.69</td>
<td>1985</td>
<td>Waste apples (Huonville, Tasmania)</td>
</tr>
<tr>
<td>nAdT.144</td>
<td>1985</td>
<td>Waste apples (Huonville, Tasmania)</td>
</tr>
<tr>
<td>nAdCH.105</td>
<td>1984</td>
<td>Decaying bananas (Coffs Harbour, New South Wales)</td>
</tr>
<tr>
<td>nAdCH.649</td>
<td>1986</td>
<td>Decaying fruit (Coffs Harbour, New South Wales)</td>
</tr>
<tr>
<td>nAdC.75</td>
<td>1988</td>
<td>Grape must (Chateau Tahbilk Winery, Victoria)</td>
</tr>
</tbody>
</table>

and in larvae. Prior to use, adult flies from each null-allele line were tested by exposure to 5% pentenol vapor to check that the line was not contaminated with normal activity Adh alleles. Two control Adh alleles encoding normal levels of ADH, Adh<sup>AC8</sup> and Adh<sup>AC5</sup>, were also used; both had been extracted from the Cygnet population in 1983. Genomic DNA was extracted from adults, either by the cesium chloride method described elsewhere (Gibson and Wilks 1989) or by the methods described by Chia et al. (1985) and Agrotis (1990). The restriction maps of some of these null alleles were obtained using endonucleases which recognized six bases (Jiang et al. 1988). Adh<sup>nAChT</sup> was analyzed in more detail by using four-base cutters and the methods described by Agrotis (1990). Two regions of genomic DNA (see below) were amplified by the polymerase chain reaction (PCR) (Saiki et al. 1988) using the same 5' primer but with two different 3' primers. The first pair of primers, JG 14 (5'-AAGCGAGGTTTCTGTATGTGTCACC-3') and JG 21 (5'-GGATTTGATTGCTTGCAGCTC-3'), were designed to amplify a 360-bp fragment of the Adh gene from nucleotides 647 to 1006 [nucleotide numbering follows that of Kreitman (1983)], which included all of exon 1 and intron 2 (fig. 1). The second pair of primers was JG 14 and JG 48 (5'-GTGGACATCGATCGATCGCTC-3'), and these were chosen to amplify a 1,135-bp fragment from nucleotides 647 to 1781, which included all three Adh exons together with introns 2 and 3 (fig. 1). Approximately 50 ng of genomic DNA was amplified using 2.5 units of Taq DNA polymerase in 100 μl of reaction mixture containing the two primers (1.0 μM each), dNTPs (each 200 μM each), and 10 μl of buffer [500 mM KCl, 100 mM TrisCl, pH 8.3, 15 mM MgCl₂, and 0.1% (w/v) gelatin]. Each PCR cycle, repeated 35 times, consisted of 30 s for denaturation of DNA at 94°C, 1 min for annealing at 60°C, and 1 min for primer extension at 72°C, with 7 min at 72°C following the final cycle. A Perkin Elmer—Cetus thermal cycler was used to control the temperatures and the ramp times. For dot-blot tests, 2 μl of PCR-amplified DNA was applied to dry nitrocellulose (Schleicher and Schuell), which was then air-dried for 10 min, denatured in a solution of 0.8 M NaCl and 0.4 M NaOH for 20 min, neutralized in a solution of 0.5 M Tris and 1.5 M NaCl (pH 7.4) for 3 min, blotted dry on Whatman 3MM paper, and baked at 80°C for ~2 h. The baked nitrocellulose filter was sealed in a plastic bag with prehybridization mixture, [3 × saline sodium citrate (SSC) containing 0.2% polyvinylpyrrolidone, 0.2% Ficoll-400, 0.2% hovine serum albumin, 0.1% sodium pyrophosphate, 0.1% sodium dodecyl sulfate (SDS), 1
Adh Null Alleles in Drosophila Populations

FIG. 1.—Structure of Adh gene of Drosophila melanogaster. The boxes denote the adult leader and the exons; the untranslated regions are hatched. The positions of two sets of primers used in the PCR to amplify genomic DNA, as well as the expected sizes of the amplified products, are indicated at (a) primers JG14 and JG21 and (b) primers JG14 and JG48. The arrow indicates the position of the oligonucleotide specific to the insertions in intron 2 of Adh\(^{nAH52}\).

mM ethylenediaminetetraacetate disodium, and 30 μg denatured herring sperm DNA/ml) and incubated for 1 h at 53°C. An 18-mer oligonucleotide that perfectly matched the sequence of Adh\(^{nAH52}\) (nucleotides 880–897) in which there are eight extra bases (in italics), 5'TCTTCAGATGCCCAGCAT-3', was labeled with γ-\(^{32}\)P-ATP (Bresatec), added to the plastic bag, and hybridized overnight at 53°C. The filter was washed at room temperature for 1 h, with the wash solution (2 X SSC containing 0.1% sodium pyrophosphate and 0.1% SDS) changed every 15 min and then was washed for 1 min at 53°C. The filter was blotted twice on Whatman 3MM paper, air-dried for ~3 hrs, was wrapped in plastic film, and was exposed to Kodak film (XR-P-1) at -70°C for ≤2 h. The EcoRI fragments containing the Adh transcription unit from each of the two null alleles Adh\(^{nACH105}\) and Adh\(^{nACH449}\) were cloned in lambda Zap II (Stratagene), and the p Bluescript phagemid was excised. The DNA sequences of both strands of Adh\(^{nACH105}\) and Adh\(^{nACH449}\) were determined from nucleotides -64 to 1802 by using the dideoxy chain-termination method (Sanger et al. 1977) with 10 synthetic oligonucleotide sequencing primers. Synthetic oligonucleotides were prepared on a 380B Applied Biosystems DNA synthesizer. The nucleotide sequence data reported for Adh\(^{nACH105}\) are in the EMBL, GenBank, and DDBJ Nucleotide Sequence Databases, under accession number M73256.

Results

Adh Null Alleles from Tasmania

The molecular structures of Adh null alleles extracted from four Tasmanian populations of Drosophila melanogaster were first compared with each other. It had previously been shown that, compared with the standard sequence (Kreitman 1983), the DNA sequence of Adh\(^{nAH52}\) and Adh\(^{nACH14}\) has eight extra bases in intron 2 (Gibson
and Wilks 1989; Freeth et al. 1990). PCR amplification of DNA from the region between nucleotides 647 and 1006 is expected to produce a 368-bp fragment in null alleles that have a structure similar to that of Adh\(^{nAH52}\) (fig. 1). Fragments of the expected size were produced using the primer pair JG14 and JG21 with genomic DNA from Adh\(^{nAH52}\), Adh\(^{nAT33}\), Adh\(^{nAC14}\), and the control normal activity Adh alleles (fig. 2). The null-allele-specific oligonucleotide probe, which contained the insertions known to be present in Adh\(^{nAH52}\), hybridized to amplified DNA from the three Tasmanian null alleles but not to that from the control alleles (fig. 2).

Six other Adh null alleles, extracted between 1983 and 1985 from four different geographically separate Tasmanian populations, were also tested using PCR-amplified DNA and the null-allele-specific oligonucleotide probe, and each of them gave a positive hybridization (fig. 2). These results demonstrate that the Adh null alleles investigated from natural populations in Tasmania all have, in intron 2, the same molecular structure shown to be the cause of the loss of ADH activity in Adht\(^{nAH52}\) and Adht\(^{nAC14}\) (Gibson and Wilks 1989; Freeth et al. 1990). Sequence data (Gibson and Wilks 1989) have shown that this type of null allele is probably derived from Adh\(^F\).

Adh Null Alleles from Mainland Australia

Three null alleles that had been isolated from mainland Australian natural populations were available for analysis and comparison with the structure of the Tasmanian

![Image](https://via.placeholder.com/150)

**FIG. 2.**—In panel a, the upper part of the figure shows genomic DNA PCR amplified using the primer pair JG14/JG21, and the lower sections show the results of probing the amplified DNA with the 18 mer oligonucleotide specific to the region of the two insertions in Adh\(^{nAH52}\). The DNA size marker is Spp1 digested with EcoRI. In panel b, the conditions are as in panel a, except that the DNA was amplified using the primers JG14 and JG48. These results were derived from four separate experiments. In each experiment there was a lane with the DNA size marker and lanes with positive and negative controls. To avoid repetition, these extra lanes are not shown. Although the hybridization intensity varied between experiments, within an experiment the intensities of controls and test samples were consistent.
null allele. To test whether they also had the insertions in intron 2, we used the same pair of primers (JG14 and JG21) in the PCR to amplify the region spanning intron 2. The amplification was successful for the null allele (Adh\textsuperscript{naC\textsubscript{H105}}) extracted from the Chateau Tahbilk population but not for the two null alleles (Adh\textsuperscript{naC\textsubscript{H449}} and Adh\textsuperscript{naC\textsubscript{H449}}) extracted from the Coffs Harbour population (fig. 2). However, the insertion-specific oligonucleotide probe did not hybridize to the PCR-amplified DNA from Adh\textsuperscript{naC\textsubscript{H105}}.

For these three null alleles, the 5' primer (JG14) was next used with a different 20-mer 3' primer (JG48), to amplify a 1,135-bp region (see fig. 1). When this second pair of primers was used, a fragment of the expected size was produced from genomic DNA of Adh\textsuperscript{naC\textsubscript{H105}} (and Adh\textsuperscript{naC\textsubscript{H449}} used as a control), but both Adh\textsuperscript{naC\textsubscript{H449}} and Adh\textsuperscript{naC\textsubscript{H449}} gave a smaller fragment of \( \sim 697 \) bp; the null-allele-specific oligonucleotide probe also failed to hybridize to any of these fragments, except that in the control Adh\textsuperscript{naC\textsubscript{H105}} (fig. 2).

These data show that the three Adh null alleles extracted from mainland Australian populations are molecularly distinct from the Tasmanian null alleles. A detailed restriction map of the 3-kb region surrounding the Adh transcription unit in Adh\textsuperscript{naC\textsubscript{H105}} was made using 4-bp-cutting restriction endonucleases. This analysis revealed an insertion of \( \sim 320 \) bp in intron 1, in the region between nucleotides 392 and 423. There was also a deletion of \( \sim 30 \) bp between nucleotides 2349 and 2516, 3' to the translated region of Adh. The analysis also suggests that this null is derived from an Adh\textsuperscript{S} allele. First, the insert \( \nabla \) at position 447 (see table 2), which Kreitmann (1983) found in three of the five Adh\textsuperscript{F} alleles he sequenced, was not present in Adh\textsuperscript{naC\textsubscript{H105}}. Second, at two sites, restriction variants that are in strong linkage association with Adh\textsuperscript{S} alleles (Kreitman 1983; Kreitman and Aguade 1986; Agrotis 1990) were present. One of these is the absence of the Ddel site at positions 1518 to 1522 (corresponding to a C at 1518 in table 2), and the other is the presence of the Ddel site at positions 1527 to 1531, corresponding to a T at 1527 (table 2).

Similar four-base cutter analyses of Adh\textsuperscript{naC\textsubscript{H105}} and Adh\textsuperscript{naC\textsubscript{H449}} (data not shown) were consonant with the data obtained by the PCR, in showing that in both genes there was a deletion of \( \sim 400 \) bp in a region that encompassed exon 2 (and consequently the binding site for primer JG21). To define these deletions and test whether they were molecularly the same, EcoRI fragments containing the Adh transcription units from Adh\textsuperscript{naC\textsubscript{H105}} and Adh\textsuperscript{naC\textsubscript{H449}} were cloned in lambda Zap II (Stratagene), and in each allele 1,866 nucleotides, spanning the Adh locus, were sequenced in each strand.

The DNA sequence data confirmed that there was a deletion of 438 bp, which included most of exon 2, in both of the Coffs Harbour null alleles. The nucleotide sequences of the two nulls were identical. The exact breakpoints of the deletion cannot be deduced because in both null alleles the sequence from nucleotide 897 is GGCTCCCTGG, whereas in the consensus sequence it is

\[
\begin{array}{c}
897 \\
\text{GGCTCCATG} \\
\text{CTCCCTGG} \\
\end{array}
\]

and hence the breakpoints could follow 898 and 1336, or 899 and 1337, or 900 and 1338, or 901 and 1339.

The other nucleotide changes that were present in the sequences of the Coffs Harbour null alleles, compared with the consensus sequences of Adh\textsuperscript{S} and Adh\textsuperscript{F}, have
### Table 2
Comparison of Consensus Nucleotide Sequences of $Adh^S$ and $Adh^F$ (from Kreitman 1983) with Four $Adh$ Null Alleles Extracted from Australian Populations

| Allele  | $-3\rightarrow-2-1$ | 107 | 113 | 293 | 304 | 423 | $\n$ | 516 | $\n$ | 816 | $\n$ | 966 | $\Delta^d$ | 1443 | 1490 | 1518 | 1527 | 1557 | 1596 | 1693 | $\n$ | $\n$
|---------|---------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| $Adh^S$ | CCG | C | C | G | A | G | G | A | ... | C | ... | T | ... | G | ... | C | A | C | T | A | G | A | ... | ...
| $Adh^F$ | TGC | T | G | C | A | G | T | C | A | Δ | G | Δ | G | ... | G | ... | G | C | T | C | C | G | C | G | C | 4 | ...
| $Adh^{MH32}$ | TGC | T | G | C | A | G | T | C | T | Δ | C | ... | T | Δ | G | ... | G | C | T | C | C | G | C | G | C | NS | NS
| $Adh^{MH105}$ | TGC | T | G | C | A | G | T | C | T | Δ | C | ... | T | Δ | T | ... | G | C | T | C | C | G | C | G | C | NS | NS
| $Adh^{MH105}$ | TGC | T | G | C | A | G | T | C | T | Δ | C | ... | T | Δ | T | ... | Deleted | Δ | G | C | T | C | C | A | C | 5 | ▼

*a* Inverted delta denotes presence of insertion at 447/477 containing a 6-bp direct repeat tandemly reiterated four times.

*b* Inverted delta denotes presence of 37-bp insertion.

*c* Inverted delta denotes presence of net 8-bp insertion.

*d* Delta denotes presence of 438-bp deletion (see text).

*e* Denotes presence of A→C polymorphism responsible for Lys→Thr difference between $Adh^F$ and $Adh^S$.

*f* Inverted delta denotes presence of variable number of A'. NS = no sequence data.

*g* Inverted delta denotes presence of extra T following 1732. NS = no sequence data.
null alleles in Drosophila populations have been previously reported in normal activity alleles (Kreitman 1983), with the exceptions of the change of A423 to T (which is also present in Adh<sup>AH52</sup> and Adh<sup>AC14</sup>) and the insertion of an extra T between 1732 and 1733 (table 2). Comparisons with the consensus sequences of Adh<sup>F</sup> and Adh<sup>S</sup> show that the Coffs Harbour null alleles are more similar to Adh<sup>F</sup> than they are to Adh<sup>S</sup> (table 2). Both null alleles had the sequence ACG at the sites (1489 to 1491) coding for the polymorphic Lys→Thr substitution, indicating that they were both derived from Adh<sup>F</sup>.

Discussion

These analyses have shown that the Adh null alleles extracted over a period of 3 years from Tasmanian natural populations of Drosophila melanogaster all have, in intron 2, the same insertions that have been shown to disrupt normal mRNA processing and to cause a lack of ADH activity (Gibson and Wilks 1989; Freeth et al. 1990). Thus the Tasmanian Adh null alleles are probably identical by descent.

Neither the data available on the frequencies of this Adh null allele nor its geographical distribution provide any clues as to where it first occurred. The null allele has been found in natural populations in the north and south of Tasmania, although its highest frequency is in one of the southern populations. The null is maintained at least in some of the populations over a number of generations, and the observed population frequencies have been close to polymorphic values, reaching nearly 4% in Huonville. Given these frequencies, migration between populations may be sufficient to disperse the allele. Various fruits, particularly apples and grapes, are transported between the northern and southern regions of the island and no doubt allow D. melanogaster to migrate between populations in the two main fruit-growing areas. Waste from apple juice manufacture is used to supplement cattle fodder in the autumn and winter months, and it is known that apple waste from the processing plant at Cygnet provided the habitat for D. melanogaster at a dairy farm in Huonville.

The population data suggest that there is only weak selection, if any, against the heterozygotes between the null and normal activity Adh alleles (Freeth and Gibson 1985). The relationship between fitness and the Adh genotypes is controversial (Van Delden 1982; Zera et al. 1983; Gibson and Wilks 1988), but the frequencies of Adh null alleles in Tasmanian populations are in agreement with the view that the differences in ADH activity may not be important for ethanol metabolism (Middleton and Kacser 1983); null heterozygotes may have a sufficient level of ADH. Early results from laboratory experiments show that Adh<sup>AH52</sup> can be maintained on normal culture media for >30 generations at polymorphic frequencies in competition with Adh<sup>F</sup>, although null homozygotes were eliminated in three generations on media supplemented with 6% ethanol (J. B. Gibson, unpublished data). This provides evidence of the strong selective disadvantage of homozygotes, which have not been detected in the natural populations sampled. Ethanol levels in decaying apples, in which D. melanogaster were breeding, have been shown (Gibson et al. 1981) to be in the range 0.06–0.39 (% v/v), although higher levels occur in other habitats.

The Tasmanian null allele has not yet been found in mainland populations, although there is the opportunity for it to spread there with the migration of D. melanogaster on transported fruit and vegetables. The null alleles that have been detected in mainland populations occur at lower frequencies. The two Adh null alleles from Coffs Harbour were isolated 2 years apart, from different banana plantations. The DNA sequences of these two alleles were identical in the region of the Adh transcription...
unit, and this suggests that they were most likely derived from a common ancestral mutation. The 438-bp deletion removed most of exon 2 and is clearly responsible for the loss of ADH activity in these two null alleles. The Coffs Harbour data provide further evidence that Adh null alleles can be maintained over several years in natural populations, again suggesting that null heterozygotes are not at a great selective disadvantage compared with Adh genotypes that have normal levels of activity.

The Adh null allele found at Chateau Tahbilk, ~1,000 km south of Coffs Harbour, represents the third form of Adh null allele so far found in natural populations, as it does not contain the structural changes shown to be present in the Tasmanian and Coffs Harbour nulls. Adh\textsuperscript{nACT} was detected in 1988 among a sample of 40 alleles that were being investigated for restriction-endonuclease variation, and its population frequency is unknown. The allele was not found in samples, taken from the same population, of 50 alleles in 1986 and of 46 alleles in 1987, and it is not yet known whether the null will persist in the population. The cause of the loss of activity in Adh\textsuperscript{nACT} cannot be deduced, with certainty, from the available data. It may be that the insertion in intron 1 disrupts transcription, but it remains possible that the cause of the loss of activity in adults and larvae is some other nucleotide change(s), present elsewhere in the gene—and this would only be revealed by DNA sequencing. Nevertheless, it is clear that Adh\textsuperscript{nACT} represents a third type of Adh null-activity allele in Australian populations.

None of the naturally occurring Adh null alleles has the same structure as do any of the artificially induced null alleles which have been studied in D. melanogaster (O'Donnell et al. 1977; Aaron 1979; Chia et al. 1985, 1987; Kelly et al. 1985; Martin et al. 1985; Batzer et al. 1988), although some, induced by formaldehyde, have deletions in the Adh coding region (Benyajati et al. 1982, 1983a, 1983b; Le et al. 1990). There is nothing in the molecular structures of either the Tasmanian or the Coffs Harbour Adh null allele to suggest the mechanism by which either of them occurred. Since there was no sign, in the DNA sequence, of the characteristic footprints left behind by some known forms of transposable elements (Finnegan and Fawcett 1985), we have discounted (Gibson and Wilks 1989) the possibility that such elements caused the lesion in the Tasmanian null allele. Similarly, in the sequence of the Coffs Harbour null, there is no evidence of this kind that would explain the origin of the deletion. The insertion in Adh\textsuperscript{nACT} has not yet been sequenced, but it is possible that it is related to one of the classes of transposable elements found in D. melanogaster populations (Finnegan and Fawcett 1985).

Finally, the geographically widespread occurrence of Adh null-activity alleles provides further evidence that alleles that do not encode a protein can be maintained in natural populations. While it is unlikely that a null allele would replace an active allele, it is possible that a null could contribute to the population variance in the molecular landscape of the Adh transcription unit, or, if the null persists and accumulates further mutations, it may acquire a new function. It could be relevant in this context that the Adh transcription unit is probably located within an intron of the gene “outsread” (Chia et al. 1985) and that structural alterations to Adh, such as those found in the null alleles, therefore may affect transcription of this region. In this way, Adh null alleles might contribute to fitness differentials associated with another gene—and hence might be of evolutionary significance.
Acknowledgments

We thank Anh Cao and Darryl Reed for assistance with DNA sequencing, Garry Mayo for making the oligonucleotides, and Cathie Stewart-Moore for preparing the manuscript.

LITERATURE CITED


GIBSON, J. B., and A. V. WILKS. 1988. The alcohol dehydrogenase polymorphism of Drosophila
*melanogaster* in relation to environmental ethanol, ethanol tolerance and alcohol dehydrogenase activity. *Heredity* **60**:403–414.


Masatoshi Nei, reviewing editor

Received March 1, 1991; revision received July 16, 1991

Accepted July 23, 1991
The *Drosophila subobscura* Adh Genomic Region Contains Valuable Evolutionary Markers¹

Gemma Marfany and Roser González-Duarte
Departament de Genètica, Universitat de Barcelona

We have sequenced 4 kb of the genomic region comprising the *Adh* (Alcohol dehydrogenase) gene of *Drosophila subobscura*. In agreement with other species which belong to the same subgenus, two structural genes, *Adh* and *Adh-dup*, are contained in this region. The main features of these two genes of *D. subobscura* have been inferred from the sequence data and compared with the homologous region of *D. ambigua* and *D. pseudoobscura*. *Drosophila subobscura* *Adh* and *Adh-dup* differ from those of *D. ambigua* at a corrected estimation of 10.1% and 12.5%, respectively, while from those of *D. pseudoobscura* they differ by 9.5% and 8.1%, respectively. Our data suggest that *Adh* and *Adh-dup* are evolving independently, showing a species-specific pattern. Moreover, particular features of some regions of these genes make them valuable evolutionary hallmarks. For instance, replacement substitutions in the third exon of *Adh* may indicate the branching of the *melanogaster-obscura* groups, whereas replacement substitutions in the third exon of the *Adh-dup* could be used to assess speciation within the *obscura* group.

Introduction

Phylogenetic relationships in *Drosophila* have been the subject of much research to illustrate evolutionary trends among species and to determine the dynamics of the different groups and subgroups. Genetic distances among species have been evaluated through the analysis of morphological, chromosomal, and biochemical traits. A wealth of information has been produced, and many phylogenetic relationships have thus been established.

The evolution of the *obscura* group of *Drosophila* has been thoroughly reviewed at morphological, cytological, and biochemical levels (Buzzati-Traverso and Scossiroli 1955; Lakovaara and Saura 1982; Krimbas and Loukas 1984; Steinemann et al. 1984; Loukas et al. 1986; Hernández et al. 1988), and electrophoretic comparisons have also been drawn for the different subgroups (Cabrera et al. 1983; Loukas et al. 1984, 1986). Lately, relationships between species in the *obscura* subgroup have been examined through the analysis of mitochondrial DNA (Latorre et al. 1988; González et al. 1990) and scDNA divergence (Goddard et al. 1990).

Comparative analysis at the genomic DNA level of one species of the *obscura* subgroup, *D. pseudoobscura*, has provided valuable information concerning evolutionary rates of individual genes (Schaeffer and Aquadro 1987) as well as on the relationships among the members of a multigene family (Brown et al. 1990). The

1. Key words: *Adh, Adh-dup, Drosophila subobscura*, gene evolution, nucleotide substitution rate, phylogenetic relationships.

Address for correspondence and reprints: Roser González-Duarte, Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Avda. Diagonal, 645, Barcelona, Spain (08028).

© 1992 by The University of Chicago. All rights reserved.
0737-4038/92/0902-0006$02.00

261
aim of the present study is to establish the evolutionary pattern of two genes contained in the Adh genomic region of D. subobscura and to make interspecific comparisons in the obscura subgroup: D. pseudoobscura (Schaeffer and Aquadro 1987), and D. ambigua (Marfany and González-Duarte 1991).

**Material and Methods**

**Fly Stock**

The source of DNA was a stock of flies of Drosophila subobscura strain H27, with standard chromosomal arrangements, provided by R. de Frutos, Department of Genetics, University of Valencia, Spain.

**Preparation of Genomic DNA and Construction of Libraries**

Total genomic DNA was isolated using the guanidine isothiocyanate method initially described for RNA extraction (Chirgwin et al. 1979), with minor modifications. High-molecular-weight total DNA was partially digested with MboI, then fractionated by sucrose gradient to obtain DNA fragments of 15–20 kb. The library was constructed by cloning this DNA into the BamHI site of an EMBL4 phage vector. Between 150,000 and 200,000 recombinant plaques were screened for sequences homologous to the D. melanogaster Adh gene. A 2.7-kb HindIII–EcoRI restriction SAC1 fragment (Goldberg 1980) containing the complete Adh gene and the immediately adjacent 3′ region was used as a probe. It was labeled either by nick-translation or by random-hexamer priming with [α-32P]dCTP (NEN-DuPont) and hybridized to phage DNA on nitrocellulose filters (Hybond-C, Amersham) in 46% formamide at 42°C overnight in the presence of 10% Na dextran sulfate. Nonspecifically hybridized probe was removed with one wash in 2 × SSC/0.1% SDS at room temperature for 10 min, then two washes in 2 × SSC/0.1% SDS at 65°C for 10 min each, and finally two further washes in 1 × SSC/0.1% SDS at 65°C for 10 min each (1 × SSC = 0.15 M NaCl/0.015 M sodium citrate, pH 7.5). After the washings, filters were autoradiographed, and the DNA from positive recombinant phages was isolated according to the method of Maniatis et al. (1982, pp. 76–85 and 371–373).

**Restriction Analysis of Positive Clones**

Positive clones were characterized by restriction mapping with the restriction enzymes BamHI, EcoRI, HindIII, PvuII, PstI, Sall, and XbaI (Boehringer Mannheim), by using single and double digests (Maniatis et al. 1982, pp. 354–376). Adh was located on the restriction map by transferring digested and electrophoresed DNA to nitrocellulose membranes. Labeling of Adh probe, hybridization conditions, washes, and autoradiography were performed as for plaque hybridization.

**Nucleotide Sequence Analysis**

Clones for sequencing were obtained by restriction digestion with suitable enzymes. These restriction fragments were subcloned into Bluescript +KS, +SK phasmids (Stratagene). Several series of nested deletions were generated using the Henikoff (1984) method to sequence large regions. Sequencing was performed on either single- or double-stranded DNA by using suitable primers according to the dideoxy method of Sanger (Sanger et al. 1977), with (α-35S)dATP (NEN-DuPont) and modified T7 phage polymerase (Sequenase from USB or Sequencing Kit from Pharmacia-LKB). After electrophoresis on TBE buffer, sequencing gels were fixed, dried, and autoradiographed for 72 h. Each nucleotide was sequenced at least three times. Sequences
were read and aligned with the 4.185-kb sequence of *D. ambigua* (Marfany and González-Duarte 1991) and the 3.535-kb sequence of *D. pseudoobscura* (Schaeffer and Aquadro 1987), as they are the only two members of the *obscura* group whose *Adh* gene has been sequenced. Alignments were determined by the Sequence Analysis Software Package of the Genetics Computer Group of the University of Wisconsin (Devereux et al. 1984). Unrooted trees were created by using the FITCH program of the PHYLIP Package (J. Felsenstein, University of Washington) based on the least-squares distance method (Fitch and Margoliash 1967). This program was chosen because it does not assume synchronous growth of branches, thus allowing differences in evolutionary rates between different species. In addition, the program DNABOOT of the PHYLIP Package (J. Felsenstein, University of Washington) was used in the calculation of bootstraps for the trees based on total coding sequence data. This program does not allow one to work with only silent or replacement positions.

**Results**

**Clone Characteristics**

Different positive clones were isolated after the screening of the library. Some of them carried the functional *Adh* gene, whereas the remainder shared the conventional features of retrosequences (also called retroposons) bearing the *Adh* coding region (authors' unpublished data). Restriction analysis of the functional *Adh* clones revealed polymorphic restriction sites possibly due to insertion/deletion events (data not shown). The cytological location of sequences homologous to *Adh* on polytene chromosomes in *Drosophila subobscura* has been reported elsewhere (Visa et al. 1991).

**Sequence Comparison of *Adh* Coding Regions**

A total of 3,980 bp of the genomic region containing *Adh* and *Adh-dup* was sequenced in *D. subobscura* (fig. 1) and then was aligned with the homologous region of *D. ambigua* (Marfany and González-Duarte 1991) and *D. pseudoobscura* (Schaeffer and Aquadro 1987).

The positions of the *D. subobscura Adh* adult and larval promoters, the proposed leaders, the transcriptional start sites, and the polyadenylation signal were determined by homology with the sequences compared. Length differences were observed in the adult and larval leaders of the three species. *Drosophila subobscura* leaders resembled *D. ambigua* not only in length but also in their position, although coding-sequence similarity was greater between *D. subobscura* and *D. pseudoobscura* (table 1). However, the overall structural comparison of the leaders of the three species showed that some of the regions were well conserved, while others shared only reduced similarity. Variation in length was observed in the intron regions: for instance, some insertion/deletion events were clearly detected in the adult intron, but functional sequences, such as consensus donor and acceptor splice sites, appeared to be highly conserved.

The *Adh* coding sequence aligned perfectly among the species compared, as would have been observed with the *melanogaster* group species, were it not for a deletion comprising six nucleotides corresponding to the third and fourth amino acids of the protein. This deletion constitutes a constant feature of the *obscura* group. The *Adh* coding region has accumulated 72 nucleotide differences since the divergence of *D. subobscura* and *D. ambigua* and has accumulated 68 between *D. subobscura* and *D. pseudoobscura*. These substitutions were not randomly distributed in any codon position, as it was evident that most of them affected the third nucleotide (*D. subobscura* vs. *D. ambigua*—\(\chi^2 = 74.34\); *D. subobscura* vs. *D. pseudoobscura*—\(\chi^2 = 67.19\);
Fig. 1.—Sequence of Adh region of Drosophila subobscura. Nucleotide position refers to the inferred adult transcription start site. The amino acid sequence of Adh and Adh-dup genes is presented under the nucleotide sequence; gaps in this amino acid sequence denote intron positions. Nucleotide differences in Adh and Adh-dup coding sequences between D. subobscura and D. ambigua are shown immediately above the nucleotide sequence, while differences between D. subobscura and D. pseudoobscura coding sequences are shown in the second line above. Replaced amino acids are also shown: underlining denotes replacements between D. subobscura and D. ambigua; wavy underlining denotes replacements between D. subobscura and D. pseudoobscura; double underlining denotes replacements that occurred between D. subobscura and either of the other two species; boxes denote replaced amino acids that the obscura species shares with D. mauritiana (melanogaster-group species representative).
df = 2 and \( P < 0.00001 \) in each case). The ratio of transitions to transversions was 1:1, thus deviating significantly from the random ratio 1:2 (D. subobscura vs. D. ambigua—\( \chi^2 = 4.0, df = 1 \) and \( P < 0.05 \); D. subobscura vs. D. pseudoobscura—\( \chi^2 = 15.56, df = 1 \) and \( P < 0.001 \)), as expected for a coding sequence. Nucleotide differences in the D. subobscura sequence, affecting the third-codon positions with fourfold degeneracy, led to an enrichment of the A+T content (table 2). A test for the expected random ratio 1:1 of A+T/G+C of these differences showed a significant deviation (D. subobscura vs. D. ambigua—\( \chi^2 = 6.91, df = 1 \) and \( P < 0.01 \); D. subobscura vs. D. pseudoobscura—\( \chi^2 = 6.43, df = 1 \) and \( P < 0.02 \)). As coding bias toward G+C is
clearly observed in the Adh gene of the Sophophora species analyzed to date (Starmer and Sullivan 1989). *D. subobscura* Adh stands out among them for having a lower G+C content (table 2).

Nucleotide substitutions that do not cause amino acid replacements are synonymous or silent, and thus they are expected to be more frequent than the rest. The number of potentially silent sites is different for each coding sequence, and for *D. subobscura Adh* it was 27.9% (213.1 of 765 nucleotide sites) (table 1). But, according
Drosophila subobscura Adh Region Markers

C C C C C C G C G G CA
A G G C C C C G C G
ACGATTTGATTCCACACAAGGCGATGGAAGTTGTGATGACCTACTGACTGATTGACATATCGCGATCGTTTTAAY

to our results, only 26.7% (57 nucleotides) of these effectively silent sites differed between *D. subobscura* and *D. ambigua*, and 28.2% (60 nucleotides) differed between *D. subobscura* and *D. pseudoobscura*. In addition, these differences were randomly distributed throughout the three exons (*D. subobscura* vs. *D. ambigua*—\( \chi^2 = 1.2, df = 2 \) and \( P > 0.5 \); *D. subobscura* vs. *D. pseudoobscura*—\( \chi^2 = 1.76, df = 2 \) and \( P > 0.2 \)).

Considering the potential replacement sites (551.9), we detected 15 differences (2.7%) between *D. subobscura* and *D. ambigua* and 8 differences (1.4%) between *D. subobscura* and *D. pseudoobscura*. These replacement differences were also randomly distributed across the three exons, whatever sequences were compared (*D. subobscura* vs. *D. ambigua*—\( \chi^2 = 0.46 \); *D. subobscura* vs. *D. pseudoobscura*—\( \chi^2 = 0.40 \); in both cases, \( df = 2 \) and \( P > 0.7 \)). Between *D. subobscura* and *D. ambigua* 12 amino acid replacements, 10 of them conservative, were produced by the 15 nucleotide differences, whereas between *D. subobscura* and *D. pseudoobscura* 7 amino acid replacements, all of them conservative, were involved (according to Feng et al.’s [1985] scale). None of the essential amino acids predicted for ADH (alcohol dehydrogenase) function (four glycines and one aspartic acid; Benyajati et al. 1981; Duester et al. 1986) has been replaced, and the ADH protein of the three species shared extensive structural similarity, as could have been predicted from biochemical data (Hernández et al. 1988).

Replacement substitutions deviated from random distribution across the three exons when the *obscura* species (Schaeffer and Aquadro 1987; Marfany and González-Duarte 1991) were compared with *D. mauritiana*, a representative member of the melanogaster group. This deviation was not present in comparisons among other *Drosophila* species (Sullivan et al. 1990). Amino acid replacements between *D. mauritiana* and any *obscura* subgroup species are boxed in figure 1. Most replacements were clustered in the third Adh exon.

Analysis of noncoding regions defined some conserved sequence motifs. Whether they play a role in the regulation of the expression of this gene remains to be tested.
Table 1
Comparative Evolutionary Analysis of Adh Genomic Region in Drosophila subobscura, D. ambigua, and D. pseudoobscura

<table>
<thead>
<tr>
<th>Region</th>
<th>D. subobscura</th>
<th>D. subobscura vs. D. ambigua</th>
<th>D. subobscura vs. D. pseudoobscura</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LENGTH (bp)</td>
<td>No. of Differences(^a)</td>
<td>% Differences</td>
</tr>
<tr>
<td>Adh:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5' Noncoding</td>
<td>32.0</td>
<td>3</td>
<td>9.4</td>
</tr>
<tr>
<td>Adult leader</td>
<td>76.0</td>
<td>17</td>
<td>23.0</td>
</tr>
<tr>
<td>Adult intron</td>
<td>745.0</td>
<td>192</td>
<td>25.8</td>
</tr>
<tr>
<td>Adult/larval leader</td>
<td>64.0</td>
<td>18</td>
<td>28.1</td>
</tr>
<tr>
<td>Exon 1:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silent</td>
<td>29.7</td>
<td>9</td>
<td>30.3</td>
</tr>
<tr>
<td>Replacement</td>
<td>63.3</td>
<td>1</td>
<td>1.6</td>
</tr>
<tr>
<td>Intron 1</td>
<td>68.0</td>
<td>13</td>
<td>19.1</td>
</tr>
<tr>
<td>Exon 2:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silent</td>
<td>112.7</td>
<td>33</td>
<td>29.3</td>
</tr>
<tr>
<td>Replacement</td>
<td>292.3</td>
<td>9</td>
<td>3.1</td>
</tr>
<tr>
<td>Intron 2</td>
<td>68.0</td>
<td>25</td>
<td>36.8</td>
</tr>
<tr>
<td>Exon 3:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silent</td>
<td>70.7</td>
<td>15</td>
<td>21.2</td>
</tr>
<tr>
<td>Replacement</td>
<td>196.3</td>
<td>5</td>
<td>2.5</td>
</tr>
<tr>
<td>Intergenic region</td>
<td>196.0</td>
<td>30</td>
<td>15.3</td>
</tr>
<tr>
<td>Adh-dup:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5' Noncoding</td>
<td>20.0</td>
<td>2</td>
<td>10.0</td>
</tr>
<tr>
<td>Leader</td>
<td>107.0</td>
<td>22</td>
<td>20.6</td>
</tr>
<tr>
<td>Exon 1:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silent</td>
<td>24.0</td>
<td>3</td>
<td>12.5</td>
</tr>
<tr>
<td>Replacement</td>
<td>72.0</td>
<td>1</td>
<td>1.4</td>
</tr>
<tr>
<td>Intron 1</td>
<td>265.0</td>
<td>65</td>
<td>24.5</td>
</tr>
</tbody>
</table>
Table 1 (Continued)

<table>
<thead>
<tr>
<th></th>
<th>Exon 2:</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Silent</td>
<td>108.3</td>
<td>41</td>
<td>37.9</td>
<td>52.7</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>296.7</td>
<td>2</td>
<td>0.7</td>
<td>0.7</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Replacement</td>
<td>62.0</td>
<td>14</td>
<td>22.6</td>
<td>26.9</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>Intron 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Silent</td>
<td>86.7</td>
<td>33</td>
<td>38.1</td>
<td>53.1</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>Replacement</td>
<td>252.3</td>
<td>17</td>
<td>6.7</td>
<td>7.1</td>
<td>12</td>
</tr>
<tr>
<td>Overall:</td>
<td>Noncoding</td>
<td>1,703.0</td>
<td>401</td>
<td>23.5</td>
<td>28.3</td>
<td>408</td>
</tr>
<tr>
<td></td>
<td>Adh:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Silent</td>
<td>213.1</td>
<td>57</td>
<td>26.7</td>
<td>33.1</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>Replacement</td>
<td>551.9</td>
<td>15</td>
<td>2.7</td>
<td>2.8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>765.0</td>
<td>72</td>
<td>9.4</td>
<td>10.1</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>Adh-dup:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Silent</td>
<td>219.0</td>
<td>77</td>
<td>35.2</td>
<td>47.4</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>Replacement</td>
<td>621.0</td>
<td>20</td>
<td>3.2</td>
<td>3.3</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>840.0</td>
<td>97</td>
<td>11.5</td>
<td>12.5</td>
<td>64</td>
</tr>
</tbody>
</table>

a The effective number of silent sites for a gene depends on its amino acid composition. It is calculated as the percentage of the potentially silent sites (possible substitutions that would not lead to an amino acid replacement) out of the total coding nucleotides (Holmquist et al. 1972). Insertions and deletions are oversimplified to one single event.

b Estimated as $d = -3/4 \ln(1-4p/3)$, where $p$ is the proportion of nucleotide sites that differ between two sequences (Jukes and Cantor 1969).

Because of the variable length of the Adh-dup third exon, we have used this table's figures for potentially silent and replacement sites when comparing *D. subobscura* vs. *D. ambigua* and have used 85.0 for silent and 251.0 for replacement sites when comparing *D. subobscura* vs. *D. pseudoobscura*. 
Table 2  
Comparative Analysis of Third-Codon Positions Having Fourfold Degeneracy for Adh and Adh-dup Sequences

A. Differences in Third-Codon Positions with Fourfold Degeneracy between *Drosophila subobscura* and either *D. ambigua* or *D. pseudoobscura* Sequences

<table>
<thead>
<tr>
<th>Sequence and Comparison</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
<th>Total</th>
<th>% G+C</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Adh:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>D. subobscura</em> vs. <em>D. ambigua</em></td>
<td>12</td>
<td>20</td>
<td>8</td>
<td>7</td>
<td>47</td>
<td>31.9</td>
</tr>
<tr>
<td><em>D. subobscura</em> vs. <em>D. pseudoobscura</em></td>
<td>11</td>
<td>14</td>
<td>5</td>
<td>5</td>
<td>35</td>
<td>28.6</td>
</tr>
<tr>
<td><strong>Adh-dup:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>D. subobscura</em> vs. <em>D. ambigua</em></td>
<td>20</td>
<td>10</td>
<td>13</td>
<td>5</td>
<td>48</td>
<td>37.5</td>
</tr>
<tr>
<td><em>D. subobscura</em> vs. <em>D. pseudoobscura</em></td>
<td>16</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>27</td>
<td>14.8</td>
</tr>
</tbody>
</table>

B. Base Composition of Third-Codon Positions with Fourfold Degeneracy in *D. subobscura*, *D. ambigua*, *D. pseudoobscura*, and *D. mauritiana*

<table>
<thead>
<tr>
<th>Sequence and Species</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
<th>Total</th>
<th>% G+C</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Adh:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>D. subobscura</em></td>
<td>15</td>
<td>29</td>
<td>55</td>
<td>35</td>
<td>134</td>
<td>67.2</td>
</tr>
<tr>
<td><em>D. ambigua</em></td>
<td>4</td>
<td>16</td>
<td>68</td>
<td>47</td>
<td>135</td>
<td>85.2</td>
</tr>
<tr>
<td><em>D. pseudoobscura</em></td>
<td>5</td>
<td>25</td>
<td>70</td>
<td>37</td>
<td>137</td>
<td>78.1</td>
</tr>
<tr>
<td><em>D. mauritiana</em></td>
<td>8</td>
<td>19</td>
<td>76</td>
<td>35</td>
<td>138</td>
<td>80.4</td>
</tr>
<tr>
<td><strong>Adh-dup:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>D. subobscura</em></td>
<td>29</td>
<td>19</td>
<td>48</td>
<td>30</td>
<td>126</td>
<td>61.9</td>
</tr>
<tr>
<td><em>D. ambigua</em></td>
<td>12</td>
<td>13</td>
<td>50</td>
<td>53</td>
<td>128</td>
<td>80.5</td>
</tr>
<tr>
<td><em>D. pseudoobscura</em></td>
<td>13</td>
<td>17</td>
<td>57</td>
<td>43</td>
<td>130</td>
<td>76.9</td>
</tr>
<tr>
<td><em>D. mauritiana</em></td>
<td>22</td>
<td>19</td>
<td>34</td>
<td>39</td>
<td>114</td>
<td>64.0</td>
</tr>
</tbody>
</table>

* Data are the number of times a change to this nucleotide has occurred in *D. subobscura*.  
  d Source: Cohn and Moore (1988).

**Adh-dup (3'ORF) Sequence Analysis**

Downstream from Adh and adjacent to it, an ORF (open reading frame) sequence, the Adh-dup, already described in some Sophophora species (Schaeffer and Aquadro 1987; Cohn and Moore 1988; Marfany and González-Duarte 1991) was detected. It shares considerable structural similarity with Adh. The main features of this gene were strongly conserved in all the species analyzed, including *D. subobscura*. It contained three exons and had all the transcriptional and translational regulation signals of an expressed gene: a CCAAT and a TATA box, splicing consensus sequences, and a polyadenylation signal, all of them located in the expected positions (fig. 1). The sequence evolution of this potential gene was analyzed in a way similar to that used for Adh.

Comparison between the Adh-dup of *D. subobscura* and those of *D. ambigua* and *D. pseudoobscura* showed considerable similarity in the first two exons, but in the third exon there were considerable length differences, because of a different location of the STOP codon, which would account for the different number of amino acids in the protein: 279 for *D. subobscura*, 281 for *D. ambigua*, and 278 for *D. pseudoobscura*
Drosophila subobscura Adh
Region Markers  271

(273 amino acids for D. mauritiana and its sibling species). The C-terminal region of the 3'ORF presented increased variability which was reflected in the number of coding nucleotides, in the amino acid sequence, and in the variation of the STOP codon position.

The pattern of nucleotide substitutions resembled that of the Adh gene. The putative coding region had accumulated a total of 97 differences since the divergence of D. subobscura and D. ambigua, and there were 64 differences between D. subobscura and D. pseudoobscura. These differences were not randomly distributed among the three positions of a codon; they tended to accumulate significantly in the third position (D. subobscura vs. D. ambigua—$\chi^2 = 73.84$; D. subobscura vs. D. pseudoobscura—$\chi^2 = 36.55$; in each case, df = 2 and $P < 0.0001$). The ratio of transitions to transversions was close to 1:1, as was found when the Adh gene was analyzed; it deviated significantly from the random ratio of 1:2 (D. subobscura vs. D. ambigua—$\chi^2 = 18.02$; D. subobscura vs. D. pseudoobscura—$\chi^2 = 17.37$; in both cases, df = 1 and $P < 0.0001$). Nucleotide differences of D. subobscura Adh-dup which affect the third codon position with fourfold degeneracy increased the A+T content of the coding sequence so that, compared with those of other species, this gene showed the lowest codon bias (table 2).

For this gene in D. subobscura (table 1), 26.1% (219.0) of the 840 coding nucleotides were effectively silent. Because of the variable length of the third exon, the number of coding nucleotides varied with species. So, when D. subobscura was compared with D. pseudoobscura, the number of effectively silent sites was 217.3 (26.0%) of 837 coding nucleotides. The number of effectively silent-site differences was higher for D. subobscura versus D. ambigua (77 nucleotides, or 35.2% of the total number of effectively silent sites) than for D. subobscura versus D. pseudoobscura (48 nucleotides, or 22.1%). Silent-site substitutions did not significantly deviate from random distribution among the three exons, either when the comparison was D. subobscura versus D. ambigua ($\chi^2 = 3.93$) or when it was D. subobscura versus D. pseudoobscura ($\chi^2 = 3.65$) in both cases, df = 2 and $P > 0.1$). Conversely, replacement substitutions clearly showed a significant deviation, whatever species was compared (D. subobscura vs. D. ambigua—$\chi^2 = 16.41$, df = 2 and $P < 0.001$; D. subobscura vs. D. pseudoobscura—$\chi^2 = 8.31$, df = 2 and $P < 0.02$). In particular, the first and second exons seemed to present a considerable constraint for replacement substitutions, while the third exon showed a great accumulation.

Alignment of the protein sequences of the Adh-dup showed 15 replaced amino acids between D. subobscura and D. ambigua and showed 13 replaced amino acids between D. subobscura and D. pseudoobscura. Most of these replacements were conservative. In contrast to the results for the Adh sequence, the Adh-dup amino acid replacements between the melanogaster-group representative (D. mauritiana) and any of the obscura species (fig. 2) were distributed randomly across the three exons ($\chi^2 = 0.02$, df = 2 and $P > 0.99$).

Discussion

Molecular data are needed to determine evolutionary pathways and phylogenetic relationships among Drosophila species. Our aim has been to obtain more information on the genome evolution of three species of the obscura group: D. subobscura, D. ambigua, and D. pseudoobscura. To this end, a 4-kb region containing the Adh gene and the Adh-dup sequence has been analyzed in detail. A very recent report cited studies which demonstrate that Adh-dup is transcribed (Kreitman and Hudson 1991).
Furthermore, its great sequence conservation in the *Sophophora* species, where it was originally described (Schaeffer and Aquadro 1987; Cohn and Moore 1988; Marfany and González-Duarte 1991), together with its many silent nucleotide substitutions, suggest that the *Adh-dup* could be a functional gene.

Nucleotide substitutions in exon and intron sites among either the *Adh* gene or the *Adh-dup* of *D. subobscura*, *D. ambigua*, and *D. pseudoobscura* show the expected pattern. Coding regions considered as a whole are more conserved than are noncoding regions, although when silent substitutions are considered they are found to have occurred with a higher frequency than have substitutions in noncoding regions (table 1). These results should be viewed cautiously, as deletions/inversions in noncoding regions are difficult to evaluate and could introduce undetectable errors. When the *Adh* coding sequences of *D. subobscura* and *D. ambigua* are considered, these two species appear to have diverged more than have *D. subobscura* and *D. pseudoobscura*. This trend is even more noticeable when the *Adh-dup* sequences are analyzed. Nevertheless, in noncoding regions the reverse seems to be true, and, as already mentioned, insertion/deletion events could introduce errors.

The accumulation of nucleotide substitutions in the third-codon position of *Adh* and *Adh-dup* as well as the deviated ratio of transitions to transversions, is consistent with the high frequency of silent substitutions. Substitutions in the third position of a codon, particularly when they are transitions, are frequently silent. If we assume that transversions would occur twice as often as do transitions and that selection removes most amino acid replacements, then the final frequency of transitions will be greater than that of transversions.

Nucleotide differences affecting the third-codon position in *D. subobscura* (data not shown) led to an enrichment of A+T base composition. This trend is clearly detected when the third-codon position with fourfold degeneracy is analyzed (table
2). A recent analysis on third-codon positions has been reported for all the *Drosophila Adh* sequences determined to date (Starmer and Sullivan 1989); the G+C value for all *Sophophora* species is ~80%. *Drosophila subobscura*, in contrast, showed a decreased G+C value of 68.7%, which is much closer to that of the distantly related *repleta* group than to that of its close relatives in its own group. That this also happens in other genes is supported by our data with the *Adh-dup* sequence, and so it may be a reasonably common feature of the *D. subobscura* genome. Nevertheless, it has been proposed that this tendency would not be favored by selection, as synonymous substitutions giving rise to A and T seem to be selected against (Ticher and Graur 1989). Whether this peculiarity in *D. subobscura* is a consequence of convergent evolution, genetic drift, or shifted codon usage because of different tRNA availability remains to be elucidated.

Analysis of the frequency, type, and distribution of substitutions shows that *Adh* and *Adh-dup* are evolving with different patterns, in agreement with other reports on several genes (Martinez-Cruzado et al. 1988; Sharp and Li 1989). All types of substitutions in the *Adh* coding sequence, as well as silent nucleotide substitutions in the *Adh-dup*, were randomly distributed among the three exons, whatever sequence was compared. Conversely, replacement substitutions in the *Adh-dup* sequence showed a strong deviation from random distribution, and most of them were clustered in the third exon. When *D. mauritiana* (or any *melanogaster*-group species) is compared with the *obscura* species, nonrandom replacement substitutions are mostly located in the third exon of the *Adh* gene, whereas in *Adh-dup* all types of substitutions are randomly distributed (Marfany and González-Duarte 1991). This result is also supported at the protein level, as revealed by cross-reactivity assays with one monoclonal antibody specific for the C-terminal domain of *D. melanogaster* ADH: positive reaction is observed with all the *melanogaster*-group species, while ADH from the *obscura* species is not detectable (J. Fibla and R. González-Duarte, unpublished data).

All these results taken together suggest that selective constraints differ in these two genes, which are located in the same genomic region. Furthermore, these selective constraints have varied, over time, to produce a specific evolutionary pattern, as deduced by the differences in the distribution and type of nucleotide substitutions among the species. The third exon of *Adh* is the region of the two genes that has nonrandomly accumulated the most differences during the divergence of the *melanogaster* and *obscura* groups, whereas within the *obscura*-group radiation the third exon of the *Adh-dup* stands out as being the most affected. Some attempts have been made to explain differences in the frequency and distribution of substitutions (Perler et al. 1980; Gillespie 1984; Palumbi 1989) and to illustrate the fact that a particular sequence can undergo variable substitution rates during evolution. Indeed, at specific periods in the course of speciation, some genomic regions could accumulate more replacement substitutions than do others, but the rate of amino acid replacement must have a limit if functionality is to be preserved. If this does not happen, new functions would appear, as is the case when gene duplication occurs. These sequences can be particularly useful to evaluate evolutionary relationships in definite periods of time and constitute outstanding candidates for evolutionary hallmarks (Perler et al. 1980). In a similar way, replacement substitutions in the third exon of *Adh* could constitute a satisfactory marker for the *melanogaster-obscura* radiation, while these replacements in the third exon of *Adh-dup* could be used as a marker within the *obscura*-group speciation.

*Adh-dup* has great structural similarity to *Adh* (table 2), as has already been observed. This similarity is reflected at the nucleotide level, 48.2% identity, significantly
deviating from the expected random value of 25% ($\chi^2 = 287.1$, df = 1 and $P < 0.0001$) and at the amino acid level, with 36.6% identity, versus a random value of 6.1% ($\chi^2 = 412.7$, df = 1 and $P < 0.0001$). As more data are made available, it becomes increasingly likely that both genes have diverged from an ancient precursor. Estimates for this duplication event give values of ~130 Myr, if one assumes that the obscura group undergoes a divergence rate of 0.4% base substitutions/Myr (Caccone and Powell 1990). This estimate is only an approximation, but it makes it plausible that Adh-dup was present in many, if not all, Drosophila radiations (Schaeffer and Aquadro 1987; R. Albalat and R. González-Duarte, unpublished data). Gene duplication events are not rare in the evolution of Drosophila. They involve not only the Adh region (Oakeshott et al. 1982; Batterham et al. 1984; Fisher and Maniatis 1985; Atkinson et al. 1988) but also other genes (Gemmill et al. 1985; Levy et al. 1985; Bewley et al. 1989; Takano et al. 1989; Brown et al. 1990) and yield a wealth of new raw material for selection or neutral drift.

In summary, the essential features of the D. subobscura Adh region are (a) a lower nucleotide substitution rate for Adh-dup than for Adh, in the comparison of D. subobscura with D. pseudoobscura, in clear contrast to what is observed in other species comparisons; (b) a higher percentage of A+T content in the third-codon position, which points to Adh and Adh-dup as being less codon biased than in other species and; (c) a higher similarity of the Adh and Adh-dup coding sequences with those in D. pseudoobscura, while noncoding regions are more similar to those in D. ambiguа. These differences should be viewed cautiously, as they may not be statistically significant; and evolutionary trees based on them are thus open to question (fig. 2).

Our data supply more information for the phylogenetic location of D. ambigua, a paleoarctic species with a controversial position. The trees based on total substitutions locate D. subobscura nearer to D. pseudoobscura than to D. ambiguа, a result which is in accordance with previous reports (Goddard et al. 1990). Indeed, bootstrap values seem to support a closer relationship for D. pseudoobscura and D. subobscura when the Adh coding sequence data are used, although this is not that obvious when the Adh-dup coding sequence is considered. On the other hand, the use of statistically significant interspecific differences overcomes the ambiguities generated in overall sequence comparisons. Replacement substitutions in the Adh-dup gene are not randomly distributed along the coding sequence, and that is why they could be particularly valuable in establishing phylogenetic relationships in our species. When these substitutions are considered, D. subobscura appears nearer to D. ambiguа than to D. pseudoobscura (fig. 2), a finding in agreement with the results obtained with mitochondrial DNA (González et al. 1990). At the same time, D. ambiguа appears to have diverged faster within the obscura subgroup, and this would explain the D. ambiguа position obtained by scDNA divergence analysis (Goddard et al. 1990). Our contribution, which uses the Adh-dup replacement substitutions as an evolutionary marker, illustrates that a rapidly diverging species, such as D. ambiguа could distort its phylogeny, which would make it appear more distantly related to other species than it really is.

The molecular analysis of the Adh region of D. subobscura has supplied relevant information on the evolutionary features of this species. From our data, it can be inferred that (1) the Adh and the Adh-dup genes show different selective constraints, (2) the evolution of the Adh and the Adh-dup genes is species specific, and (3) a particular gene or even a specific coding region may be under variable evolutionary constraints, yielding different nucleotide substitution rates and different distribution
of these substitutions. These sequences are excellent genetic markers for evolutionary studies.

**Sequence Availability**

Accession numbers for sequence data in GenBank/EMBL Data Library are as follows: *Drosophila subobscura*, M55545; and *D. ambigua*, X54813.

**Acknowledgments**

We are grateful to Dr. A. Prevosti for encouragement and valuable discussions and to R. Rycroft for revising the English version of the manuscript. This research was supported by Dirección General de Investigación Científica y Técnica (DGICYT) grant PB88-0188 and by CIRIT (Generalitat de Catalunya) grants AR 87 and AR 88. G. M. is in receipt of an FPI fellowship.

**LITERATURE CITED**


BRIAN CHARLESWORTH, reviewing editor

Received July 1, 1991; revision received July 29, 1991

Accepted September 25, 1991
Traditional fruiting body–based classification of ascomycetes has been under attack for 2 decades. Fruiting-body types can converge, and few researchers now assume that either the closed fruiting bodies (cleistothecia) characterizing the class Plectomycetes or the flask-shaped fruiting bodies (perithecia) characterizing the class Pyrenomycetes are stable, unifying characters. Unless we identify characters uniting major ascomycete groups, orders of ascomycetes remain narrowly defined, and supraordinal classification is impossible. We sequenced both strands of 18s rDNA from nine ascomycete fungi, adding three sequences from GenBank into our analysis. The phylogeny, inferred from 162 informative sites in 1,700 bp of DNA sequence data and using yeast as an outgroup, divided the fungi into two groups correlating well both with fruiting-body type and with the traditional classes Plectomycetes and Pyrenomycetes. Each group received strong statistical support. Genera producing cleistothecia, such as Talaromyces (with a Penicillium asexual state) and the human pathogen Ajellomyces capsulatus (causing histoplasmosis), fall within the plectomycete group. Plectomycetes also includes Eremascus albus and the bee pathogen Ascosphaera apis, although both lack typical fruiting bodies. The Dutch elm disease fungus groups with pyrenomycetes such as Neurospora, in spite of its confusing mixture of class-level characters.

Introduction

The Ascomycotina, including almost 40% of known fungus species, is the largest subdivision of fungi (Hawksworth et al. 1983). Ascomycetes produce spores in an ascus, or sac, and traditional ascomycete classes have been defined both by the form of the fruiting bodies bearing the asci and by the type and arrangement of the asci. Members of the class Plectomycetes have cleistothecia, closed fruiting bodies characterized by an irregular distribution of asci and ascospore release following disintegration or deliquescence of the ascus wall (fig. 1) (Fennell 1973). In the past, the class Plectomycetes included the organisms with Penicillium or Aspergillus asexual states, as well as the fungi causing ringworm and histoplasmosis. Pyrenomycetes included Neurospora and other ascomycetes forming flask-shaped perithecia with a single layer of asci and forcible discharge of ascospores from the ascus (fig. 1) (Müller and von Arx 1973).

The problem with traditional classification is that fruiting-body characters can converge (Cain 1972; Malloch 1981). A species normally producing flasklike fruiting bodies can be induced to form closed fruiting bodies under certain environmental
Fig. 1.—Two classes of ascomycetes corresponding to phylogenetic groups inferred from rDNA sequence data. The percentages are the frequencies with which a given branch appeared in 500 bootstrap replications. Branches that appeared in $\geq95\%$ of the replications are strongly supported by the data set. Branches with frequencies $<75\%$ have been reduced to polychotomies. The two most parsimonious trees generated by the Branch and Bound option in PAUP require 462 changes. The above tree is identical to one generated from a consensus of the 25 most parsimonious Branch and Bound trees with lengths under 464. (The graphic depicting Pyrenomycetes is reproduced, with permission from the publisher, from Fungal Spores: Their Liberation and Dispersal, by C. T. Ingold. Oxford: Oxford University Press, 1971. The graphic depicting Plectomycetes is reproduced, with permission from publisher, from Introduction to Fungi, by J. Webster. New York: Cambridge University Press and Academic Press, 1980.)

conditions (von Arx 1973). Some genera of fungi with flasklike fruiting bodies are closely related to genera of fungi with closed fruiting bodies (von Arx 1973). Some species, such as *Ophiostoma ulmi* (the Dutch Elm disease fungus), have both the fruiting-body type characteristic of the Pyrenomycetes and the ascus arrangement characteristic of the Plectomycetes. Without characters uniting major ascomycete groups, orders of ascomycetes must remain narrowly defined, and supraordinal classification is impossible (Eriksson 1982; Hawksworth et al. 1983). At present, Plectomycetes and Pyrenomycetes and other classes are falling into disuse (Hawksworth et al. 1983) (fig. 2). Classless taxonomic systems, while prudent, offer no help to the molecular biologist wanting to know whether the human pathogenic fungi are (1) more closely related to *Aspergillus* than to *Neurospora* and therefore (2) more likely to undergo homologous than nonhomologous transformation (Fincham 1989).

If fruiting-body characters indicate common descent, we would expect that the genera we studied would be divided into two groups corresponding to traditional
1971
6 classes
25 orders
HEMIASCOMYCETES
Endomycetales
Protomycetales
Taphrinales
PLECTOMYCETES
Ascospheareales
Eurotales
Microascales
Erysiphales
Meliolales
PYRENOMYCETES
Hypocreales
Sphaeriales
Clavicipitales
Corynellaes
Coronophorales
DISCOMYCETES
Phacidiales
Helotiales
Ostropales
Pezizales
Tuberales
LABOULBENIOMYCETES
Laboulbeniales
LOCULOASCOMYCETES
Myriangiales
Pezizales
Sordariales
Hysteriales
Sphegidales
Capnodiales
Microthyriales

1983
0 classes
37 orders
Arthoniales
Saccasphaerales
Calicibales
Clavicipitales
Corynellaes
Cystidiales
Diaporthales
Diatrycales
Dothidiales
Elaphomyctales
Endomycetales
Erysiphales
Eurotales
Graphidiales
Gyalectales
Gymnoascales
Helotiales
Hypocreales
Laboulbeniales
Lecanidiales
Lecanorales
Microascales
Ophidostomatales
Ostropales
Peltigerales
Pezizales
Polystigmatales
Pyrenulales
Rhytismatales
Sordariales
Sphathulosporales
Sphaeriales
Taphrinales
Teloschistales
Verrucariales

FIG. 2.—The Dictionary of the Fungi, a standard reference for fungal terminology, placed ascomycete orders into classes in 1973 but dropped the classes from the 1983 edition. The number of orders within the Ascomycotina increased from 25 in the 1973 edition to 37 in the 1983 edition. By eliminating classes and defining orders narrowly, the dictionary minimizes the chances of uniting unrelated, morphologically convergent organisms but provides little information about higher-level relationships. Superscript numbers show the classification of fungal genera included in the present study: 1 = Saccharomyces; 2 = Thermoascus; 3 = Byssomycetes; 4 = Talaromyces; 5 = Monascus; 6 = Ajellomyces; 7 = Eremascus; 8 = Ascosphaera; 9 = Neurospora; 10 = Chaetomium; 11 = Leucostoma; and 12 = Ophiostoma. [Columns are reproduced, with permission from the publisher, from Hawksworth et al. (1983).]

Ascomycete classes and correlating with fruiting-body characters. If fruiting-body characters arose convergently, we would anticipate that phylogeny inferred from sequence might correlate with other morphological characters. Either way, sequence characters
had the potential to suggest which morphological characters usually indicate phylogenetic relationships leading to a more cohesive system of ascomycete taxonomy.

Material and Methods

We sequenced the nuclear 18S rDNA of nine ascomycetes in six orders (Hawksworth et al. 1983) and added sequence from Neurospora crassa (GenBank NEURRNAS), Ajellomyces capsulatus (GenBank X58572), and Saccharomyces cerevisiae (GenBank YSCRGEA) into our analysis (fig. 1). We amplified the rDNA subunit from miniprepped DNA by using primers NS1 and NS8 and 30 cycles (each cycle = 2 min at 97°C, 1 min at 48°C, and 45 s at 72°C, with a 4 s/cycle extension at 72°C) of the polymerase chain reaction (PCR) (Lee and Taylor 1990; White et al. 1990). We sequenced single-stranded template from asymmetric amplification of double-stranded PCR template by using primer pairs including NS1-8 (White et al. 1990) and NS19-22 (A. Gargas, personal communication), with the primer in excess at 0.5 μm and the limiting primer at 0.025 μm (White et al. 1990). Only one strand was sequenced near primers NS 1 and NS 8, and in an ~100-nucleotide-long region near NS 5. Otherwise, both strands were sequenced. We aligned the sequences visually and excluded ambiguously aligned sites from our analysis. Phylogenetic trees with identical topologies were generated using either the maximum-parsimony method, PAUP 3.01 (Camin and Sokal 1965; Swofford 1989), or the distance neighbor-joining (Saitou and Nei 1987) method. Sequences have been deposited in GenBank, and fungal strains followed by GenBank accession codes are Ascospbaira apis UCB 78-018 (M83264), Byssochlamys nivea FRR 2205 (M83256), Chaetomium elatum UCB 81-063 (M83257), Eremascus albus UCB 50-026 (M83258), Leucostoma persoonii LP8 Gerry Adams personal collection (M83259), Monascus purpureus ATCC 16365 (M83260), Ophiostoma ulmi ATCC 32437 (M83261), Talaromyces flavus var. macrospora FRR 2386 (M83262), and Thermoascus crustaceus FRR 1328 (M83263) (UCB = University of California, Berkeley collection; FRR = Food Research Laboratory, North Ryde, New South Wales; and ATCC = American Type Culture Collection). Alignment is available on request.

Results and Discussion

From ~1,700 bp of sequence per fungus, 1,628 sites were well aligned for all 12 fungi. Out of the 302 variable sites, the 162 phylogenetically informative sites were the basis for inferring phylogenetic relationships by using maximum-parsimony methods (Camin and Sokal 1965; Swofford 1989). Eleven ascomycete genera in seven orders (Hawksworth et al. 1983) fall into two groups, with Saccharomyces cerevisiae, a 12th fungus, as an outgroup in a parsimony-based phylogenetic tree from sequence of 18S nuclear ribosomal RNA (rDNA). The groups correspond to traditional ascomycete classes Plectomycetes and Pyrenomycetes. The first group, the Plectomycetes, includes organisms with cleistothecial fruiting bodies, as well as both Ascospbaira apis (causing chalk brood disease of bees), which has a vesicle rather than a hyphal cleistothecial wall, and Eremascus albus, which lacks a fruiting body. The second group, the Pyrenomycetes, includes Neurospora crassa, with typical pyrenomycete characters, and the Dutch elm disease fungus Ophiostoma ulmi, with a pyrenomycete-like fruiting body containing a plectomycete-like distribution of asci. We did a bootstrap analysis and found that branches leading to Plectomycetes and Pyrenomycetes were supported at the 100% level. Levels >95% indicate strong statistical support for branches
At least 42 nucleotide changes showing no homoplasy occurred on the branches leading either to Plectomycetes or to Pyrenomycetes, and 18 of these were unambiguously assignable to each branch. The yeast *S. cerevisiae* was used as an outgroup, on the basis of (1) its distance from the other taxa included in the present study, (2) morphological considerations, and (3) the results of preliminary study comparing the sequence of the small nuclear subunit of rDNA of chytridiomycetes, basidiomycetes, and ascomycetes (Bowman et al. 1992).

The plectomycetes, with morphologically simple reproductive structures, have been viewed either as primitive or as a heterogeneous assemblage of reduced, unrelated ascomycetes. We found that morphologically diverse representatives of Plectomycetes together form a distinct monophyletic group (fig. 1). Within the Plectomycetes, saprobic *Penicillium* or *Penicillium*-like asexual states, forming dry chains of asexual spores from specialized cells (phialids), grouped together (family Trichochomaceae: *Talaromyces, Thermoascus, and Byssochlamys*) (Malloch and Cain 1972) along with, unexpectedly, *Monascus purpureus* (a fungus exuding red-purple pigment used in coloring some Asian foods). *Monascus purpureus* had been placed in its own family both because of its unusual sex organs (gametangia and cleistothecia) and because it lacks the phialidic asexual spores characterizing the Trichochomaceae (Cole and Kendrick 1968).

*Ajellomyces capsulatus* (the human pathogen causing histoplasmosis) and the three human pathogens most closely related to it (B. Bowman, personal communication) are also plectomycetes, making typical cleistothecia if reproducing sexually.

*Ascosphaera apis*, causing chalk brood disease of bees, and *Eremascus albus* have been taxonomic puzzles. They have been classified with yeasts, because they lack a hyphal fruiting body (Harrold 1950; Fennell 1973; Kreger-von Rij 1973), or in Plectomycetes, because they are mycelial, form hyphal gametangia, and have eight ascospores in each ascus (Spiltoir and Olive 1955; von Arx 1981, pp. 88 and 138). On the basis of rDNA sequence, the two are plectomycetes, demonstrating the strength of rDNA sequence data in placing organisms when diagnostic cleistothecial morphology is lacking.

The pyrenomycetes that we examined varied both in ascus structure and in presence of sterile tissue surrounding the perithecia, but all are similar in form, and, with the exception of the Dutch elm disease fungus, all are generally considered to be related (von Arx 1981, pp. 19 and 150–178; Barr 1990). The Dutch elm disease fungus *Ophiostoma ulmi* was originally considered to be a member of Pyrenomycetes because of its dark-colored, flask-shaped fruiting bodies (Müller and von Arx 1973; von Arx 1981, p. 150). However, like plectomycetes, its asci are not organized into a single layer in the fruiting body, and ascospores are released when the ascus wall deliquesces, rather than through a forcible discharge mechanism (Nannfeldt 1932; Benny and Kimbrough 1980). When sequences are compared, the Dutch elm disease fungus groups with other pyrenomycetes rather than with plectomycetes or with the yeast *Saccharomyces cerevisiae*.

We have found that sequence data support a simple but controversial taxonomic hypothesis—i.e., that ascomycetes from different orders but with similar fruiting-body characters belong in the same class. Not all fungi in a class have all characters of the class, and sequence data can indicate relationships when some morphological characters are atypical. We demonstrated that the fruiting-body type, rather than ascus type and arrangement, of the Dutch elm disease fungus was most consistent with its class-level relationships. As the nucleic acid sequences of more ascomycetes become available,
the boundaries of the Plectomycetes, Pyrenomycetes, and other classes will be defined.
With equal certainty, support for the Pyrenomycetes and Plectomycetes will remain strong, judging from the robust statistical support for branches grouping the diverse organisms included in the present study.

Acknowledgments

We thank D. R. Reynolds, B. Bowman, and T. Bruns for critical comments and thank A. Gargas for unpublished primers. This work was supported in part by NIH grant R01 AI28545.

LITERATURE CITED


WALTER M. FITCH, reviewing editor

Received June 7, 1991; revision received July 29, 1991

Accepted September 3, 1991
Molecular Evolution of the Fungi: Relationship of the Basidiomycetes, Ascomycetes, and Chytridiomycetes

Barbara H. Bowman,* John W. Taylor,† Alan G. Brownlee,‡ Jean Lee.§ Shi-Da Lu,* and Thomas J. White*

*Roche Molecular Systems; †Department of Plant Biology, University of California, Berkeley; ‡Division of Animal Production, CSIRO; and §Cetus Corporation

Establishing the phylogeny of fungi and protists often has proved difficult owing to the simple morphologies and convergent characters in these organisms. We used DNA sequences of nuclear small-subunit ribosomal RNA genes to determine phylogenetic relationships among three major classes of organisms considered to be fungi—Basidiomycetes, Ascomycetes and Chytridiomycetes—and to assess the taxonomic position of Neocallimastix, an economically important anaerobic rumen microorganism whose classification is controversial. The Basidiomycetes and Ascomycetes, two classes of nonflagellated fungi, are the most closely related taxa. Chytridiomycetes, though bearing flagella, group with these higher fungi rather than with the protists. Neocallimastix, a eukaryote lacking mitochondria and variously classified as a protist or as a fungus, shows closest molecular affinities with the Chytridiomycete fungi in the order Spizellomycetales.

Introduction

The fungi, historically considered to be plants because of their form and apparent lack of locomotion, have been recognized as a kingdom in their own right, a classification championed by Whittaker (1969). Two of the classes of fungi that lack flagella—Basidiomycetes (mushrooms, “shelf fungi,” and their allies) and Ascomycetes (mostly molds and yeasts)—are universally allied by morphological characters (Tehler 1988). Also traditionally classified as fungi are three classes of flagellated organisms—Chytridiomycetes ("chytrids"), Oomycetes (water molds), and Hyphochytriomycetes. Of these, only the chytrids share with the nonflagellated classes their cell-wall polymers (Bartnicki-Garcia 1970) and lysine synthetic pathway (Vogel 1964).

Phylogenetic analysis based on the ~120-base 5S ribosomal RNA (rRNA) suggested that the nonflagellated fungi do not form a natural (monophyletic) group (Hori and Osawa 1987); indeed, in that analysis, green plants branched between the two fungal classes. Other researchers (Van de Peer et al. 1990) have found these nonflagellated classes to form a monophyletic group, also based on 5S rRNA. However, because of the small amount of information in the 5S molecule, neither of these results is likely to be strongly supported (Halanych 1991; Steele et al. 1991). Hendriks et al. (1991), using 18S rDNA, found that the nonflagellated classes were not monophyletic when a limited number of sequence positions were used in the analysis but that these classes became so when a less conservative subset was used. To determine whether

1. Key words: molecular evolution, 18S ribosomal RNA, rumen fungi, chytrid, basidiomycete.

Address for correspondence and reprints: Barbara H. Bowman, Roche Molecular Systems, 1145 Atlantic Avenue, Suite 100, Alameda, California 94501.

© 1992 by The University of Chicago. All rights reserved.
0737-4038/92/0902-0008$02.00

285
these conflicting observations can be resolved with any confidence, we have used our sequence of nuclear small-subunit ribosomal DNA (18S rDNA) from the basidiomycete shelf fungus *Spongipels unicolor*, comparing it to sequences already known from Ascomycetes and other organisms.

Phylogenetic placement of the class Chytridiomycetes is a matter of current debate. These organisms are claimed for the protists on the basis of possession of flagella (Margulis and Schwartz 1988, pp. 77, 140–141, and 153) and are claimed for the fungi on the basis of characters that they share with the nonflagellated classes. Förster et al. (1990), using 18S rDNA sequences, placed the chytrid *Blastocladiella emersonii* as a sister group to the ascomycetes, but the degree of support for this conclusion was not addressed. We present two new sequences for 18S rDNA from members of the Chytridiomycetes and determine with a high degree of confidence their relationship to the nonflagellated fungi.

*Neocallimastix* and related anaerobic gut microbes also have controversial classifications. Although these agriculturally important, cellulolytic microbes share characters with Chytridiomycetes, many of their characteristics are unique, i.e., multiple flagella (Braune 1913), genomic G+C content <20% (Brownlee 1989), lack of mitochondria (Heath et al. 1983), strict anaerobiosis (Orpin 1975), and hydrogen-generating hydrogenosomes (Yarlett et al. 1986). The zoospores of these gut microbes share ultrastructural features with one order of Chytridiomycetes (Spizellomycetales) (Barr 1980; Heath et al. 1983) and share developmental features with another (Blastocladiales) (Wubah et al. 1991); but intense interest in these microbes has resulted in many new descriptions, and their classification is anything but settled (Barr et al. 1989; Breton et al. 1990; Ho et al. 1990). To determine the phylogenetic placement of *Neocallimastix*, we sequenced its nuclear 18S rDNA, for comparison with our chytrid sequences and other known sequences.

We therefore address three questions that are well suited to molecular evolutionary investigation: (1) Are the Basidiomycetes, Ascomycetes, and Chytridiomycetes part of a monophyletic fungal lineage? (2) Are the anaerobic gut microorganisms members of the Chytridiomycetes? (3) Are gut chytrids most closely related to the order Spizellomycetales?

**Material and Methods**

*Neocallimastix* species, culture LM-2, was obtained from the Anaerobic Culture Collection, CSIRO, Prospect, New South Wales. *Spizellomyces acuminatus*, strain Barr 62A, was a gift from Donald Barr. *Chytridium confervae* 81-1 was obtained from the U.C. Microgarden (University of California, Berkeley). Total DNA isolated from *C. confervae* (Lee and Taylor 1990) and *Neocallimastix* (Brownlee 1988) and pSK2 plasmid DNA from *Spongipels unicolor* (Kwok et al. 1986) was amplified using conditions which minimize the errors introduced by the *Taq* DNA polymerase (<1 error/5 kb after 35 cycles) (Gelfand and White 1990) with modified versions of primers NS1 and NS8 (White et al. 1990): SL21—CCGAATTCGTCATATGCTTGTCT; SL27—CCAGCATACCGAGCTACGACT. Amplified DNA was purified by precipitation, treated with Klenow fragment, purified by agarose electrophoresis, and blunt-end ligated into the *HincII* site of the pUC18 vector. Transformed *Escherichia coli* strain DH5 was screened with the probe NS2 (White et al. 1990), and clones were sequenced from both strands (Sequenase protocol; U.S. Biochemical). Sequences obtained from the cloned polymerase chain reaction (PCR) products of *Neocallimastix*, *C. confervae*, and *Spongipels unicolor* were verified against partial sequences obtained
from direct sequencing of 800–1,100 bases each from single-stranded DNA, asymmetrically amplified (Kwok et al. 1986; Gyllensten and Erlich 1988) either from 1–10 ng of genomic DNA or from the cloned Spongipelis unicolor rDNA repeat.

PCR reactions on Spizellomyces acuminatus total DNA were performed using primers NS1 (White et al. 1990) and NS24 (AAACCTTGTACGACTTTTA) (A. Gargas, personal communication), one of which was biotinylated in each of two reciprocal reactions. Spizellomyces acuminatus DNA was diluted 1:100 with autoclaved distilled water from the extracted stock DNA. Fifty microliters of diluted DNA was amplified in a 100-µl reaction containing 1 x GeneAmp Buffer (Perkin Elmer–Cetus); 62.5 µM each dGTP, dATP, dTTP, and dCTP; 5% glycerol (Smith et al. 1990); 50 pmol of each primer; and 2.5 units AmpliTaq (Perkin Elmer–Cetus). The reaction mixture was heated to 95°C for 5 min and then was subjected to 40 cycles, each of 40 s at 95°C, 25 s at 50°C, and 3 min at 72°C; these were followed by a single 10-min extension at 72°C. Products were stored at 4°C briefly before single-stranded-DNA preparation.

Spizellomyces acuminatus single-stranded 18S rDNA for sequencing was prepared using a streptavidin agarose (SA) technique based on that of Mitchell and Merrill (1989), with the variation that columns were eliminated, and the entire procedure was done in microcentrifuge tubes. All procedures were performed at room temperature. After each step, the mixture was spun briefly in a microcentrifuge to pellet the SA beads, and the supernatant was removed with an Eppendorf micropipetter. SA beads (Bethesda Research Laboratories 5942SA) were rinsed five times with storage buffer (20 mM Tris, 200 mM NaCl, 1 mM ethylenediaminetetraacetate, pH 7.6) before use. The slurry of SA and storage buffer (~1:1 [v:v]) was stored at 4°C. Approximately 90 µl of double-stranded, biotinylated PCR product was captured on 200-µl SA slurry in a 2-ml microcentrifuge tube (Sarstedt 72,689) by agitation on a Labquake rotator (Labindustries, Berkeley, Calif.) for 30–40 min. After being washed twice with 500 µl storage buffer, the double-stranded DNA (still attached to SA) was denatured by rinsing the beads twice, for 6 min each time, with 150 µl freshly prepared 0.2 M NaOH, and the nonbiotinylated single strand of DNA was recovered in the supernatants. The combined supernatant containing the ssDNA was neutralized with 200 µl of 5 M ammonium acetate, pH 6.8, and was desalted and concentrated to a volume of 35–60 µl by using three 2-ml washes with autoclaved distilled water on Centricon-100 (Amicon 4212). The resulting ~40 µl of single-stranded DNA solution, 7.5 µl was sequenced using standard Sequenase 2.0 protocol (U.S. Biochemical), with the exception that the labeling mix was diluted 1:20 before use. Bases at 1,417 positions were sequenced on both strands of the DNA. Single-stranded sequence was accepted when the data were clear, were read identically by two investigators who had no knowledge of the other’s reading, and were alignable with published fungal sequences.

Results and Discussion

We obtained sequences for 18S rRNA genes, complete except for an estimated 38 bases at the 5’ end and 49 bases at the 3’ end, for Neocallimastix species, the two chytrids Chytridium confervae and Spizellomyces acuminatus, and the basidiomycete shelf fungus Spongipelis unicolor (fig. 1). These could be unambiguously aligned with published sequences from the chytrid Blastocladiella emersonii (Förster et al. 1990), the ascomycete bread mold Neurospora crassa, the soybean Glycine max, and the ciliate protist Stylonychia pustulata, at 1,368 positions (underlined in fig. 1); 308 of these positions showed at least one substitution event. The maximum number of
FIG. 1.—Aligned small subunit rRNA sequences from four species. Sequence data were obtained from the anaerobic rumen fungus *Neocallimastix* (1,717 bases; GenBank accession number M59761), the chytrids *Chytridium confervae* (1,707 bases; M59758), *Spizellomyces acuminatus* (1,718 bases; M59759), and the basidiomycete shelf fungus *Spongipellis unicolor* (1,718 bases; M59760), between PCR primers NS1 and NS24. These were aligned with published sequences of the ascomycete bread mold *Neurospora crassus* (GenBank NERURNAS), the chytrid *Blusrocladiellu emersonii* (Forster et al. 1990), the ciliate protist *Stylonychiu pustulata* (GenBank SLURGSS), and the soybean *Glycine max* (GenBank SOYRGE). (These four sequences are not shown, but the complete alignment will be supplied on request.) Sequences were aligned manually, using the Eyeball Sequence Editor ESEE version 1.09 (Cabot and Beckenbach 1989) for
Molecular Evolution of Fungi 289

the IBM PC. A dot indicates a base both homologous to and identical to the corresponding base in the *Spongipelis unicolor* (reference) sequence. A capital letter indicates a base that differs from the base in the reference sequence. A lowercase letter indicates a base at a position whose homology to positions in the reference sequence is undetermined. These include areas in which the sequences are sufficiently diverged that positional homology cannot be inferred directly from sequence similarity and in which, in addition, small or large length variations are evident. When short (<12 bases) areas of substantial sequence divergence were bounded by regions of certain homology and included no length variation, these were considered alignable (i.e., positional homology was inferred—e.g., see positions 743–747 in fig. 1). Unalignable areas were omitted from all phylogenetic analyses. In some regions, a subset of the sequences was alignable, and for those sequences these positions were included in parsimony analyses. Positions used in distance analyses (neighbor joining) are underlined and include only those positions alignable in all eight species (including *Neurospora crassa*, *B. emersonii*, *Stylonychia pustulata*, and *G. max*). The boundary of a sequence that was alignable despite high sequence variation was differentiated from an adjacent unalignable area if (1) three of the four bases at the margin of the aligned region were identical to those at the same positions in at least one other sequence of certain alignment, (2) these three matching bases included the position bounding the aligned area, and (3) no length variation was involved (e.g., see *C. confervae* at positions 605–608, which is identical both to the reference sequence at positions 606 and 608 and to the well-aligned *Neocakmastix* sequence at position 605).

Pairwise substitutions in the aligned regions was 169 (12.4%) between the bread mold and the ciliate protist.

We used three different methods of phylogenetic analysis to determine evolutionary relationships and to evaluate the strength of the branching order: the neighbor-joining method (Saitou and Nei 1987), based on genetic distances; a bootstrap test of the neighbor-joining method (Whittam 1991); and the parsimony-based Winning Sites Test (Prager and Wilson 1988).

The neighbor-joining method (Saitou and Nei 1987) was used to construct, from the unambiguously aligned positions, the branching order shown in figure 2. This tree unites the two classes of nonflagellated fungi as closest relatives, in accordance with morphological and physiological criteria. The anaerobic rumen fungus *Neocallimastix* falls within a monophyletic group formed by the flagellated class Chytridiomycetes. This group is united with the nonflagellated fungal classes, to the exclusion of the protist.

Any phylogenetic tree must be considered a hypothesis; it is necessary to examine rigorously any specific inferences drawn from this tree, to see whether the data are sufficiently robust to support definitive conclusions. The neighbor-joining method usually returns the minimum-length tree (Saitou and Nei 1987; Saitou and Imanishi 1989), but this is not necessarily the correct tree. (Neighbor-joining is guaranteed to return the correct tree only if distance data satisfy the condition of being additive.) In order to evaluate the strengths of particular inferences derived from the neighbor-
FIG. 2.—Neighbor-joining tree of eight species. The computer program NJTREE (Jin and Ferguson 1990) was used to estimate the branching order and branch lengths for the eight species. Of the 1,368 alignable sequence positions, a maximum of 169 (12.4%) differed in any pairwise comparison. Uncorrected differences were used to construct the tree. Branch lengths have been rounded to the nearest whole number of substitutions, and branches are drawn proportional to their lengths. A scale is provided to show the inferred number of substitutions per sequence position. In this tree, nonflagellated fungi form a monophyletic group. The anaerobic rumen flagellate Neocallimastix falls within the monophyletic group formed by the Chytridiomycetes (flagellated fungi). Chytrids and the nonflagellated fungi form a monophyletic group to the exclusion of the protist and green plant. The two significant bootstrap results (Whittam 1991), returned after 1,000 trials, are shown in brackets. Bootstrap support for other branches was not significant at the 95% level (Felsenstein 1985): there was 93.1% support for the branch uniting Neocallimastix with Spizellomyces acuminatus, 81.3% support for the branch uniting these two with Chytridium confervae, and 93.4% support for the monophyly of the chytrids (flagellated fungi).

The Winning Sites Test determines how strongly the branching order is supported by the data, by contrasting the number of sequence positions that support alternative, unrooted tree topologies. For the simplest test, four taxa are arranged into the three possible topologically distinct, bifurcating networks (e.g., see fig. 3). Phylogenetically informative sequence positions for this test are those at which two species share one base while the two others share a different base. When a common base is shared by the two species on the same side of the network, the data can be explained by a single substitution event, occurring on the central branch between the nodes connecting the pairs of sequences. The other two topologies would require at least two events. The network requiring a single event is counted as “winning” at that sequence position. The total numbers of sequence positions favoring each of the three networks are compared, to see whether support for one topology significantly exceeds support for the alternatives (Li and Gouy 1991). Because the Winning Sites Test can be led astray by base-compositional bias among the sequences compared (Irwin and Wilson, accepted) the base compositions of the sequences analyzed were evaluated. Neocallimastix genomic DNA is known to have an unusually low percentage of G+C (18%; Brownlee...
Molecular Evolution of Fungi 291

Fig. 3.—Evaluation of relationship of Ascomycetes and Basidiomycetes. The Winning Sites Test (Prager and Wilson 1988) was used to compare support for the branching order suggested by 5S rRNA sequence comparisons and for the morphologically based branching order, which is also that returned by the neighbor-joining method using 18S rDNA sequences. A, 5S-based rooted tree (left) and corresponding unrooted network. B, One possible rooted tree that unites nonflagellated fungi in monophyletic group. The unrooted network (right) represents any rooted tree uniting these two taxa. The third possible network is also shown. Fifty sequence positions are phylogenetically informative for these four taxa, with two taxa sharing one base and the other two sharing a different base. Of these, 11 support (require only one substitution in) network A, 30 support network B, and 9 support network C. When the four-taxon method of Li and Gouy (1991) is used, this result is significant at the 0.15% level.

1989); nevertheless, the percentage of G+C in Neocallimastix 18S rDNA is 41.9% and does not differ markedly from those of the other fungi evaluated, whose percentages of G+C are 44.3%–48.2%.

Using the Winning Sites Test, we first evaluated whether the nonflagellated fungal classes are closest relatives. In addition to these fungi, green plants and ciliate protists were represented both in the present study and in the 5S rRNA study by Hori and Osawa (1987) that found the nonflagellated fungi not to be monophyletic. The 5S rRNA tree grouped the Ascomycetes (represented here by the bread mold Neurospora crassa) with green plants (the soybean G. max) and grouped the Basidiomycetes (represented here by the shelf fungus Spongipellis unicolor) with ciliate protists (Stylonychia pustulata). (This branching order, obtained using a simplified unweighted pair-group method, was not strongly supported, having substantial error bars on the closely spaced branch points.) We compared the network representing the 5S tree (fig. 3a) both with that representing the 18S neighbor-joining tree (fig. 3b) and with the third alternative. Of 50 sequence positions phylogenetically informative for these four taxa, 11 supported the 5S branching order, 30 supported the 18S branching order, and 9 supported the third alternative. This support for the unity of the nonflagellated fungal classes is significant at the 0.15% level, by the four-taxon method of Li and Gouy (1991).
Because the protists are a paraphyletic group, their relationship to the fungal classes might not be accurately represented by a single phylum. We repeated this test, using the chrysophyte alga *Ochromonas danica* (GenBank OCHAB), whose divergence from other lineages was close in time to the divergence of plant and fungal lineages (Gunderson et al. 1987). Thus *Ochromonas* stands the best chance of disrupting the monophyletic grouping of the fungi. Of 43 phylogenetically informative positions, 26 grouped the two nonflagellated fungal lineages, while only 6 and 11 positions, respectively, supported the two alternatives (support again was significant at the 0.15% level) (Li and Gouy 1991). Thus the Winning Sites Test strongly favors the branching order suggested by morphological and physiological criteria.

In addition, a bootstrapping test using the neighbor-joining algorithm (Whittam 1991) supported the association of the nonflagellated fungal classes, in 976 out of 1,000 replicates (fig. 2). The greater quantity of information in the 18S molecule (relative to 5S) can indeed be used to resolve with a high degree of confidence the branching order of these four taxa, supporting the morphological classification uniting the two classes of nonflagellated fungi.

A second inference from the neighbor-joining tree is that the Chytridiomycetes are the closest relatives of the nonflagellated fungi. Chytrids traditionally have been classified as flagellated fungi, along with the Oomycetes (water molds) and Hyphochytriomycetes. Margulis and Schwartz (1988, p. 76) show a morphologically based tree in which chytrids are not closely related to the nonflagellated fungi and are classed as protists. Recent molecular studies (Kwok et al. 1986; Förster et al. 1990) have suggested instead both that chytrids are part of a monophyletic lineage that includes the nonflagellated fungal classes and that Oomycete water molds have their closest relatives among the protists.

We used the Winning Sites Test to determine whether the proposition that chytrids shared their most recent common ancestry with nonflagellated fungi, rather than with protists, could be supported with a high degree of confidence. For this analysis, we compared the soybean 18S rDNA sequence both with those of the ciliate protist *Stylonychia pustulata*, a species from the Chytridiomycete type genus *Chytridium*, and with the nonflagellated shelf fungus *Spongipelis unicolor*. Twenty-five sequence positions supported grouping *C. confervae* with the shelf fungus, while only 12 united it with the protist, and 7 favored the third alternative, supporting, at the 0.9% level, the association of the chytrids with the nonflagellated fungi (Li and Gouy 1991). Substitution of the bread mold for the shelf fungus gave similar results, as did using the chrysophyte *O. danica* in place of the ciliate. In addition, a bootstrapped neighbor-joining test (Whittam 1991) supported (in 991 of 1,000 replicates) the branch unifying the flagellated (Chytridiomycete) and nonflagellated fungal lineages, to the exclusion of the protist and green plant (fig. 2).

A third inference tested using the Winning Sites Test is that the anaerobic rumen fungus *Neocallimastix* is part of a monophyletic fungal lineage, rather than being a protist. Because of the paraphyly of the protists, testing the simple association of *Neocallimastix* with another fungus, to the exclusion of a representative protist, is insufficient. It is necessary to demonstrate that *Neocallimastix* branches within the fungi and is not an exterior branch.

Along with *Neocallimastix*, the ciliate protist *Stylonychia pustulata*, the bread mold *Neurospora crassa*, and the chytrid *Spizellomyces acuminatus* were used to test this inference (fig. 4). If the tree uniting *Neocallimastix* with the protist is favored (fig. 4a), then *Neocallimastix* might not be a fungus. If it unites with either of the
FIG. 4.—Evaluation of placement of anaerobic rumen flagellate *Neocallimastix* in Fungi kingdom. The Winning Sites Test (Prager and Wilson 1988) was used to evaluate the strength of the inference that *Neocallimastix* belongs in the Fungi kingdom and is not a protist. Support of network a (topologically equivalent to the rooted tree to its left) would allow the conclusion that *Neocallimastix* is a protist; support for either network b or network c would place *Neocallimastix* within the Fungi kingdom. Of the positions phylogenetically informative for these taxa, 19 favor placing *Neocallimastix* within the fungi and associated with the chytrid *Spizellomyces acuminatus*, and 4 positions favor each of networks a and c. When the four-taxon method of Li and Gouy (1991) is used, this result is significant at the 0.15% level.

other species, it can with confidence be called a fungus (fig. 4b and c). Of the 27 positions phylogenetically informative for these four taxa, 19 unite *Neocallimastix* with the Chytridiomycete fungus, while only 4 place it with the protist, and 4 associate it with the bread mold (this distribution is significant at the 0.15% level) (Li and Gouy 1991). This result supports the placement of *Neocallimastix* within the fungi. Specifically, *Neocallimastix* is allied with the chytrid branch, to the exclusion of the nonflagellated fungal lineage. The more conservative bootstrap test of neighbor-joining (Whittam 1991), however, provided nonsignificant support (e.g., uniting the four flagellated fungi as a monophyletic group in only 931 of 1,000 trials.)

To evaluate with which order *Neocallimastix* is allied within the class Chytridiomycetes, we compared its sequence with those of three other fungi: (1) *B. emersonii*, representing the order Blastocladiales, which shares some developmental features with *Neocallimastix* (Wubah et al. 1991); (2) *Spizellomyces acuminatus*, representing the order Spizellomycetales, into which Heath et al. (1983) proposed placing *Neocallimastix* on the basis of zoospore ultrastructure; and (3) the shelf fungus as outgroup. The network uniting *Neocallimastix* with the order Spizellomycetales was overwhelmingly supported (22 sequence positions favoring *Spizellomyces acuminatus* vs. 4 and 3 favoring *B. emersonii* and the outgroup, respectively), verifying this aspect of the neighbor-joining tree (with support at the 0.15% level) (Li and Gouy 1991) and supporting Heath et al.'s (1983) classification based on zoospore ultrastructure. When
we substituted *C. confervae* (Order Chytridiales) for *B. emersonii* in the Winning Sites Test, the test could not exclude the possibility that *Neocallimastix* united with the order Chytridiales instead of with the Spizellomycetales. Bootstrap analysis (Whittam 1991) returned nonsignificant results, uniting *Neocallimastix* with Spizellomyces in 934 of 1,000 trials. Nevertheless, the phylogenetic position of *Neocallimastix* is as a late branch within the fungi, suggesting that its lack of mitochondria is due to secondary loss.

In summary, we have presented four new essentially full-length fungal 18S rRNA sequences from one basidiomycete and three chytrid fungi. Within the organisms represented by these sequences, our data support the association of the nonflagellated fungal classes Ascomycetes and Basidiomycetes, in keeping with morphological classification. We have further shown, with a high degree of confidence, that the flagellated class Chytridiomycetes and the nonflagellated fungal classes together form a monophyletic unit. Although the Chytridiomycetes were the first class to diverge at the base of the fungal branch, there is no strong reason to exclude them from the Fungi kingdom.

We demonstrate with the present study an instance in which molecular data have been used successfully to discern the phylogenetic status of an unusual organism whose unique biochemical properties made its classification by morphological and physiological criteria a matter of contention. The rumen flagellate *Neocallimastix* is certainly a Chytridiomycete fungus. It has been possible, with a high probability, to assign this genus to the order Spizellomycetales, to the certain exclusion of the order Blastocladiinae.

**Acknowledgments**

We thank A. Wilson, D. Barr, E. Zimmer, and N. Arnheim, and the reviewers for their comments on the manuscript; D. Spasic, L. Goda, and C. Levenson for oligonucleotide synthesis; M. A. Brow for advice on sequencing; M. Sogin for providing the *Blastocladiella emersonii* sequence in advance of publication; R. Holmquist, J. Lake, and W.-H. Li for helpful discussion; A. Gargas for the sequence of primer NS24; and J. Price for encouragement and support. This research was supported in part by NIH grant AI28545-02.

**LITERATURE CITED**


JIN, L., and J. W. H. FERGUSON. 1990. NJTREE (v2.0) (computer program). Center for Population Genetics, University of Texas Health Science Center at Houston, Houston.

KWOK, S., T. J. WHITE, and J. W. TAYLOR. 1986. Evolutionary relationships between fungi, red algae, and other simple eukaryotes inferred from total DNA hybridization to a cloned basidiomycete ribosomal DNA. Exp. Mycol. 10:196–204.


WHITTAM, T. S. 1991. PSFIND and NJBOOT. (computer programs distributed by the author.) Pennsylvania State University, University Park.


MASATOSHI NEI, reviewing editor

Received March 5, 1991; revision received July 23, 1991

Accepted July 25, 1991
The Two Similarly Expressed Genes Encoding U3 snRNA in *Schizosaccharomyces pombe* Lack Introns

David A. Selinger, Gregory L. Porter, Patrick J. Brennwald, and Jo Ann Wise
Department of Biochemistry, University of Illinois

Both genes encoding U3 small nuclear RNA (snRNA) from the budding yeast *Saccharomyces cerevisiae* were recently shown to be interrupted by introns of the type removed by the pre-mRNA splicing machinery. We previously described one of the two U3 genes from the fission yeast *Schizosaccharomyces pombe*. In the present work, the second *S. pombe* U3 coding sequence was identified, and direct RNA sequence analysis was used to show that neither the U3A nor the U3B gene from this organism contains an intervening sequence. Our data also demonstrate that, as expected, the two RNAs exhibit great primary- and secondary-structure conservation. These similarities are not likely to be the result of a recent gene duplication or conversion event, because the DNA sequences flanking the U3A and U3B genes have diverged substantially. A notable exception is a 19-bp block, centered 36 nucleotides upstream from the transcriptional start site, in which the two loci match in 15 positions; this motif may represent an RNA polymerase II upstream regulatory element, because related sequences are found preceding fission yeast U1, U2, U4, and U5 snRNA genes. The significance of a short conserved sequence just downstream of the U3A and U3B genes is unknown, as it is not found 3' to other snRNA coding sequences in *S. pombe*. The 5' one-third of U3B RNA can be folded into a dual hairpin structure, as we previously proposed for *Schizosaccharomyces pombe* U3A and for other lower eukaryotic U3 homologues. Quantitation of fission yeast U3A and U3B indicates that, in contrast to snR17A and B in *Saccharomyces cerevisiae*, these RNAs accumulate to similar levels.

Introduction

The nuclei of eukaryotic cells contain a variety of small, stable RNA molecules, among which are the uridine-rich small nuclear RNAs (U-snRNAs). Five of the major species in this class—U1, U2, and U4–U6—are located in the nucleoplasm and play roles in splicing of messenger RNA (mRNA) precursors (reviewed in Guthrie and Patterson 1988), while U3 resides in the nucleolus and participates in preribosomal RNA (pre-rRNA) processing (reviewed in Gerbi et al., accepted). Like U1–U5, U3 possesses a trimethylguanosine cap structure and is complexed with proteins. Phylogenetic comparisons have revealed four regions of extended sequence similarity, shared
by all U3 homologues, designated boxes A–D (Wise and Weiner 1980; Hughes et al. 1987). Boxes C and D have also been found in two minor human U-snRNAs—U8 and U13—that also fractionate with the nucleolus (Reddy et al. 1985; Tyc and Steitz 1989) and in the essential yeast snRNA U14 (Jarmolowski et al. 1990).

Genes encoding homologues of vertebrate U3 RNA have been cloned from a variety of organisms, including slime mold (Wise and Weiner 1980), budding yeast (Hughes et al. 1987), fission yeast (Porter et al. 1988), and tomato (Kiss and Solymosy 1990). The 3′ two-thirds of each of these can adopt a secondary structure similar to that derived from chemical and enzymatic probing of the human RNA by Parker and Steitz (1987). Our previous analysis of one of the two genes encoding U3 from Schizosaccharomyces pombe (U3A) indicated that, in contrast to human U3, the 5′ end of this RNA cannot adopt a single stable stem-loop structure. Because the Dictyostelium homologue also readily formed a dual hairpin and since, conversely, human U3 cannot assume such an alternative structure, we suggested that between higher and lower eukaryotes there may be structural divergence in this snRNA (Porter et al. 1988). This hypothesis was also based on analysis of the gene sequences for snRl7A and B, the U3 homologues from Saccharomyces cerevisiae. Recently, a reinvestigation at the RNA level revealed that the 5′ termini originally assigned to these RNAs (Hughes et al. 1987) were incorrect because, unexpectedly, both genes are interrupted by an intron between positions 14 and 15 of the mature RNA (Myslinski et al. 1990). Here we report the structure of the gene encoding Schizosaccharomyces pombe U3B, together with RNA analyses demonstrating that neither coding sequence for U3 in this organism contains an intervening sequence. Quantitation reveals that, also in contrast to the situation in budding yeast, these genes are expressed at similar levels.

Material and Methods

Construction of a Lambda-DASH Library

Schizosaccharomyces pombe genomic DNA was partially digested with Sau3AI and inserted into the BamHI site of the replacement vector Lambda-DASH (Stratagene), which requires 9–23 kb of inserted DNA to form an infectious particle. Recombinant plaques were selected by growth on a restrictive host, P2-392, that permits growth of phage lacking the red and gam genes.

Screening the Library with a Probe Derived from the U3A Gene

The genomic library was probed with a U3A subclone labeled with [α-32P]dCTP by using the random hexamer method according to the instructions provided by the manufacturer (Amersham). This plasmid contains 116 bp from the 3′ end of the U3A coding sequence, plus 61 bp of 3′ flanking DNA. Hybridization and low-stringency washing conditions were as described elsewhere (Porter et al. 1988) for probing genomic Southern blots with a T7 transcript derived from this plasmid. After rescreening and plaque purification, DNA was prepared from selected phage by the method described by Maniatis et al. (1982, pp. 371–372).

Identification of Candidate U3B Genes, Subcloning, and Sequencing

To distinguish between phage carrying the U3A locus and those that were novel isolates, the appropriate restriction-enzyme digests (see Results) were resolved on a 0.7% agarose gel and were transferred to GeneScreen-plus (NEN). Blots were hybrid-
ized with the same probe used in the library screen. A 1.3-kb EcoRI fragment derived from one of the two U3B candidates was subcloned into pTZ18U, in both orientations. Sequence analysis was carried out by the dideoxy chain-termination method using the strategy described in Results.

RNA Analysis

Protocols for enzymatic and primer-extension RNA sequence analysis were as described elsewhere (McPheeters et al. 1986; Brennwald et al. 1988). To determine the relative abundance of U3A and U3B, we used primer extension in the presence of dATP, dCTP, dTTP, and ddGTP from an oligonucleotide complementary to a region just downstream of a sequence difference between the two RNAs. Densitometry was performed on an LKB Ultroscan; the results reported are the average of four scans.

Results

Identification of a U3B Gene

Our previously published Southern blot analysis indicated that the Schizosaccharomyces pombe genome contains two copies of the U3 coding sequence (Porter et al. 1988); only one of these genes was present in the library screened in that study. To isolate the U3B gene, we probed ~40,000 phage plaques from a different genomic bank recently constructed in our laboratory (see Material and Methods). If an average insert size of 16 kb is assumed, then this represents ~40 genome equivalents. Fifty-three plaques hybridized to the probe, in reasonable agreement with the number expected for a gene present in two copies. After plaque purification, DNA was prepared from six of these and was digested with EcoRI, EcoRI/HindIII, XbaI, and XbaI/HindIII; these enzymes cleave in the vicinity of the U3A gene (Porter et al. 1988). Two of the six phage DNAs had restriction patterns that matched the U3A locus, two had patterns related to each other but not to U3A, and the remaining two were unrelated to either U3A or the other five recombinant clones. The last were apparently false positives, since no hybridizing bands were observed on Southern blots of their restricted DNAs. The two candidate clones for U3B hybridized on genomic Southern blots to bands whose sizes were consistent with earlier genomic Southern data (Porter et al. 1988) and also cross-hybridized to blots of the putative U3A clones (G. L. Porter and J. A. Wise, unpublished data). Further restriction mapping showed that the two U3B isolates contain overlapping regions of the genome, truncated at different points by the Sau3AI partial digestion. A restriction map of the insert in the smaller phage, together with that of an EcoRI fragment subcloned into pTZ18U, is shown in figure 1. Our U3B map has sites in common with a less detailed map derived by Dandekar and Tollervey (1989), suggesting that they cloned the same locus; however, their placement of the SspI and AccI sites, which fall within the region we sequenced, was significantly different.

Comparison of the U3A and U3B Primary and Secondary Structures

The putative U3B plasmid was initially sequenced using an oligonucleotide that hybridizes to nucleotides 19-33 of U3A RNA, verifying that it contained both the complement of the primer and related but divergent upstream sequences. To obtain the complete sequence of both strands of the gene and flanking DNA, we constructed
FIG. 1.—Restriction map of U3B locus and sequencing strategy. The positions of restriction sites in the 11.0-kb Lambda-DASH insert are shown on the top line. A detailed restriction map of the 1.3-kb EcoRI fragment that was inserted into pTZ18U is shown on the bottom line. The broad, black arrow indicates the U3B coding sequence, and the half-arrows below it show the sequencing strategy.

four subclones by deleting DNA extending to either the SspI or AccI site from each orientation of the 1.3-kb EcoRI fragment (see fig. 1). Sequencing was carried out according to the strategy depicted in figure 1 by using the universal primer or an oligonucleotide complementary to the U3B 3’ flanking sequence, as appropriate. Figure 2 shows the secondary structure we propose for U3B RNA, with primary sequence differences in U3A indicated. As expected, the two fission yeast U3 coding sequences are closely related: only 22 differences were found in 255 nucleotides (91% identity). snR 17A and B from Saccharomyces cerevisiae (Hughes et al. 1987; Myslinski et al. 1990) have a similar degree of identity, 92% (28 differences in 333 nucleotides), while rat U3A and U3B (Reddy 1989) are more diverged, with 28 differences in 214 nucleotides (87% identity). Boxes A–D, the regions of highest similarity among all U3 homologues (Wise and Weiner 1980; Hughes et al. 1987), are especially well conserved between Schizosaccharomyces pombe U3A and U3B; the only differences are two transitions in box B. The two RNAs differ in size by one nucleotide, because of an extra U at position 210 in U3A. Size heterogeneity was also observed between U3A and U3B RNAs of rat (Reddy 1989).

The proposed Schizosaccharomyces pombe U3B secondary structure is similar to the one we originally reported for U3A (Porter et al. 1988), with several minor modifications. First, the loop at the top of hairpin 1a has been reduced by four nucleotides, because of the addition of one G·U and one A·U pair to the stem. Second, hairpin 1b and surrounding nucleotides have been refolded to reflect the U3B sequence and the corrected U3A sequence. The new structure proposed for this region contains a stem-loop and spacer in place of a longer hairpin, which is satisfying from a phylogenetic perspective because it conforms more closely to the folding pattern we proposed earlier for the 5’ end of Dictyostelium U3 (Porter et al. 1988). Note, however, that stems 1a and 1b are not supported by the existence of compensatory base changes. The validity of hairpin 2 is supported by the exchange of G-C for A-U pairs between Schizosaccharomyces pombe U3A and U3B RNAs. Other sequence differences in hairpins 2 and 3, while not compensatory, maintain the helices. Both ends of U3A and U3B are remarkably devoid of sequence differences—none occur either between
One impetus to identify the second fission yeast U3 gene was to test the model we previously proposed for a dual hairpin near the 5' ends of lower-eukaryotic homologues of this RNA. A single stem-loop is the only stable structure available to the corresponding region of U3 from vertebrates. Because up to position 91 there are no sequence differences between fission yeast U3A and U3B RNAs, the present work does not provide additional support for the dual hairpin model. However, recently published data from other laboratories bear on this issue. First, the structure we proposed for the 5' end of *Saccharomyces cerevisiae* U3 (Porter et al. 1988) was based...
on an incorrect sequence, because of the unanticipated finding that both snR17 genes contain introns (Myslinski et al. 1990); the revised sequence can also be folded into a dual hairpin structure. Second, the 5' ends of Arabidopsis, tobacco, and tomato U3 RNAs can be folded into dual hairpins similar to those we previously proposed for lower-eukaryotic U3 (Kiss and Solymosy 1990; Marshallsay et al. 1990). Although conservation of U3 structure between plants and fungi might at first seem surprising, we note that in many phylogenetic trees the plant-animal divergence appears deeper than that between plants and fungi (e.g., see Sogin et al. 1989).

An abrupt decrease in sequence conservation is observed at the boundaries of the fission yeast U3A and U3B coding regions; however, as illustrated in figure 3, the flanking DNA does retain blocks of identity. The underlined sequence centered at −36 corresponds to a conserved element found preceding all Schizosaccharomyces pombe U-snRNA genes except for U6; we have previously proposed that this sequence serves as an RNA polymerase II transcriptional control element (Porter et al. 1990). Downstream positions +33 to +39 in U3A perfectly match positions +35 to +41 in U3B, and just upstream is a 10/12 identity; the significance of this stretch of conserved nucleotides is unclear, because other fission yeast snRNA loci lack similar sequences in this region (Small et al. 1989, and unpublished data cited therein).

U3 Genes from Schizosaccharomyces pombe Lack Introns

The introns in the budding yeast U3 genes occur between nucleotides corresponding to positions 14 and 15 of the mature RNAs (Myslinski et al. 1990). We employed a combination of direct RNA sequencing strategies to address the question of whether either or both fission yeast U3 coding sequences might also be interrupted. First, as shown in figure 4, primer-extension sequence data extending from position 80 to the 5' end demonstrate that this portion of the RNA (which is identical between U3A and U3B) is colinear with the genes (see fig. 2). Thus, the coding sequences are uninterrupted throughout this interval and, in particular, are continuous through the location of the budding yeast U3 intron. We located the 3' boundary of the genes by using enzymatic RNA sequencing of 3' end-labeled U3 RNA, as shown in figure 5. Northern analysis of total, nuclear, and anti-TMG precipitated RNA with several
different U3-specific probes (Porter et al. 1988; G. L. Porter and J. A. Wise, unpublished data) produced a single band indistinguishable in size from 7SL RNA, which is 254 nucleotides long (Brennwald et al. 1988); the resolution of these gels was not sufficient to detect the one-base difference in length between U3A and U3B. Taken together, these data demonstrate unequivocally that neither gene encoding U3 RNA in *Schizosaccharomyces pombe* contains an intron.

Fission Yeast U3A and U3B Genes Are Expressed at Similar Levels

The two genes encoding snR17, the *Saccharomyces cerevisiae* U3 homologue, are expressed at 5-10-fold different levels (Hughes et al. 1987). To determine the relative abundance of *Schizosaccharomyces pombe* U3A and U3B RNAs, we devised a primer-extension strategy (see Material and Methods) that would produce a single distinct product for each of the RNAs. The results, shown in figure 6, indicate that U3B represents 60% of the steady-state RNA and that the remaining 40% corresponds to the U3A sequence. (Quantitation was performed on a gel exposure that was ninefold-

![Fig. 4.—Primer-extension RNA sequencing of 5' end of Schizosaccharomyces pombe U3 RNA. The products of reverse-transcribing total RNA from the primer U3-PE (5' GTGCCCTCTATCATCC 3'), complementary to residues 78–97 of *S. pombe* U3 RNA, are displayed on a 10% sequencing gel. Lanes G, A, T, and C, Primer-extension RNA sequencing. The sequence indicated alongside the gel is the complement of the dyeoxynucleotide incorporated at that position and thus corresponds to the RNA sequence written 5' to 3', top to bottom. Lane PE, Products of primer extension from same oligonucleotide in absence of dyeoxynucleotides.](image-url)
fainter than the one depicted.) Thus, in contrast to the dramatic difference in abundance of the transcripts from *Saccharomyces cerevisiae* U3 genes, the *Schizosaccharomyces pombe* genes appear to contribute almost equally to the pool of RNA.

**Discussion**

The discovery by Myslinski et al. (1990) that both genes encoding snR17 in *Saccharomyces cerevisiae* are interrupted was unexpected. U3 genes isolated from a slime mold (Wise and Weiner 1980), a plant (Kiss and Solymosi 1990), and several vertebrates (Reddy 1989) are all colinear with the encoded RNAs; thus, introns are not a common feature of sequences coding for this small nuclear RNA (snRNA). We show here that neither of the two U3 genes from another yeast is interrupted. Although there is disagreement about the precise evolutionary distance between fission yeast and budding yeast, phylogenies based on either morphological or molecular markers
Fig. 6.—Quantitation of Schizosaccharomyces pombe U3A and U3B transcripts. Primer extension was carried out on total RNA from the oligonucleotide U3-BoxD (5' GTCAGAAAAACACCAGCTG 3'), complementary to nucleotides 233–250 of U3B and to nucleotides 234–251 of U3A. Lanes G, A, T, and C. Standard primer-extension RNA sequencing reactions. The sequence indicated alongside the gel is the complement of the dideoxynucleotide incorporated at that position and thus corresponds to the RNA sequence written 5' to 3', top to bottom. Lane-dG. Sequencing reaction in which only ddGTP (and no dGTP) was included. The position of the primer-extension product from each RNA is indicated.

generally place them both among the ascomycetes. The fact that fission yeast U3 genes are colinear with their products suggests that the snR17 introns may have been acquired after the Saccharomyces cerevisiae lineage diverged from other fungi. Although Schizosaccharomyces pombe U3 genes lack introns, it was the first organism in which an mRNA-like intron was discovered in an snRNA coding sequence, specifically the single-copy gene encoding the spliceosomal RNA U6 (Tani and Ohshima 1989); analysis of U6 genes from a variety of fungi including all extant Schizosaccharomyces species indicated that this intron was most likely not present in the ancestral copy (Reich and Wise 1990). It will be interesting to learn whether organisms more closely related to Saccharomyces cerevisiae possess interrupted U3 genes.

Although U3 was the first snRNA shown to be conserved across wide evolutionary distances (Wise and Weiner 1980), it is only recently that its function has been addressed experimentally. Kass et al. (1990) used immunodepletion and RNAse H–mediated cleavage of U3 in an in vitro extract to demonstrate a role for this snRNA in the first detectable preribosomal RNA processing event in mammals. This was consistent with earlier data indicating that U3 can be psoralen cross-linked to the prerRNA near the site of 5' external transcribed spacer (ETS) cleavage (Maser and Calvet
A different role for U3 is suggested by the results of Savino and Gerbi (1990), who found that RNase H–mediated depletion of the snRNA in *Xenopus* oocytes decreased cleavage at the internal transcribed spacer 1 (ITS1) processing site, causing accumulation of two precursor ribosomal RNAs (rRNAs) downstream in the pathway, 20S and 32S. These incompatible observations may reflect species differences or, more likely, may indicate that U3 does not participate directly in cleavage reactions but, rather, promotes or stabilizes a particular conformation of the pre-rRNA compatible with both processing and ribosome assembly (Savino and Gerbi 1990).

Whatever the precise role of U3 may be, its 5′ end appears to be functionally important, because the site of cross-linking to the ETS lies within or near box A of rat U3 (Stroke and Weiner 1989) and because oligonucleotides that target RNase H cleavage to box A or to the spacer separating the 5′ end from the proximal stem disrupt rRNA processing (Kass et al. 1990; Savino and Gerbi 1990). The ability of the 5′ end of U3 to fold into a dual hairpin appears to be widespread (Myalinski et al. 1990; Kiss and Solymosy 1990; Marshallsay et al. 1990)—rather than being restricted to unicellular eukaryotes, as we originally proposed (Porter et al. 1988). Since frog is the only nonmammal among the vertebrates from which U3 has been characterized (Jeppeisen et al. 1988), sequences of homologues from phylogenetically intermediate organisms would be useful to determine whether the existence of a single 5′ hairpin is a very recent adaptation. Unlike most RNA components of ribonucleoproteins, the universally conserved nucleotides near the 5′ end of U3 are not single stranded, and, moreover, as noted by Gerbi et al. (accepted), the stems are not supported by compensatory base-pair substitutions in homologues from different organisms. Because it seems likely that the role of U3 has been conserved through evolution, we favor the idea that the stable 5′ helices, whether single or double, are not required for U3 snRNP function. They might, for example, serve to prevent this portion of the RNA from engaging in aberrant folding with the remainder of the molecule prior to its assembly with proteins.

**Acknowledgments**

We are grateful to Claudia Reich for helpful comments on the manuscript. This research was supported by NIH grant 1 R01 GM38070 awarded to J.A.W.

**LITERATURE CITED**


ALAN M. WEINER, reviewing editor

Received July 15, 1991; revision received July 30, 1991

Accepted September 9, 1991
Character Analysis and the Integration of Molecular and Morphological Data in an Understanding of Sand Dollar Phylogeny

Charles R. Marshall
Department of Earth and Space Sciences, University of California

Reconciling discordant morphological and molecular phylogenies remains a problem in modern systematics. By examining conflicting DNA-hybridization and morphological phylogenies of sand dollars, I show that morphological criteria may be used to help evaluate the reliability of molecular phylogenies where they differ from morphological trees. All available criteria for assessing the reliability of DNA-hybridization phylogenies suggest that the sand dollar DNA-hybridization phylogeny is robust. Standard homology-recognition criteria are used to assess the a priori reliabilities of the morphological attributes associated with the node drawn into question by the DNA data, and it is shown that these attributes are among the least phylogenetically informative of all the morphological characters. Moreover, the questioned node has the smallest number of supporting characters, and most of these characters are associated with the food grooves, which suggests that they may be functionally correlated. Thus, on the basis of the analysis of the morphological data and given the robustness of the DNA tree, the DNA phylogeny is preferred. Further, paleobiogeographic data support the DNA tree rather than the morphological tree, and a plausible heterochronic mechanism has been proposed that may account for the homoplasious morphological evolution that must have occurred if the DNA tree is correct.

Introduction

How to integrate molecular and morphological data most effectively is one of the challenges of contemporary phylogenetics. Molecular and morphological phylogenies frequently exhibit broadly similar topologies, with disagreements restricted to the position of a few taxa. In these cases the reliability or robustness of both phylogenies is assessed (e.g., by bootstrapping), and the tree with the greatest support is favored. In some cases the molecular and morphological characters are combined in a single analysis (e.g., see de Sá and Hillis 1990), and the most parsimonious tree for the combined data, if statistically robust, is accepted.

Here I use an alternative approach to integrating discordant molecular and morphological phylogenies. The approach is based on the fact that characters established in morphological analyses represent testable hypotheses of relationship (Bryant 1989). Criteria exist for evaluating the relative informativeness or reliability of characters prior to their use in building evolutionary trees (Remane 1952), and it is the use of these criteria to establish the relative reliabilities of the characters drawn into question

1. Key words: character analysis, DNA hybridization, sand dollar phylogeny, homology recognition criteria, paleobiogeography, heterochrony.

Address for correspondence and reprints: Department of Earth and Space Sciences, University of California, Los Angeles, Los Angeles, California 90024-1567.

© 1992 by The University of Chicago. All rights reserved.
0737-4038/92/0902-0010$02.00

309
by molecular data that offers the possibility of using morphological data to help evaluate the robustness of molecular phylogenies. These homology-recognition criteria are traditionally used to establish phylogenetically informative characters in the initial stages of morphological analyses, and the process of applying the criteria is termed "character analysis" by some (e.g., Bryant 1989; Sluys 1989). As an example of the application of the second of the three major recognition criteria of Remane (1952) ("similarity between structures based on the correspondence between numerous separate features of these structures"; Sluys 1989, p. 353), architecturally very simple characters are more likely to reflect undetected homoplasies than are characters with highly intricate and complex architectures. Riedl (1978, pp. 34–36), Wiley (1981, pp. 130–131), and Sluys (1989) provide discussions of Remane's (1952) criteria for recognizing homologies.

Marshall and Swift (accepted) provide a DNA-hybridization phylogeny of sand dollars that differs from phylogenies derived from morphological traits. Here I use the general principles of character analysis to establish a priori relative weights for the morphological characters. These weights are then used to assess the extent to which the morphological data are contradicted by the DNA phylogeny. It is shown that, while the DNA-hybridization phylogeny is essentially unanticipated by the morphological data, relatively few characters—and only those with relatively high a priori likelihoods of being undetected homoplasies—are drawn into question by the DNA phylogeny. Further morphological and paleobiogeographic data are presented that support the DNA tree where it differs from the morphological cladogram.

Sand dollars were selected for the present study because of their relatively complex morphology and excellent fossil record (Marshall 1988). An analysis of the fossil record of the group is being prepared which assesses the absolute rates of single-copy DNA evolution in the group and employs the methods outlined by Marshall (1990a, 1990b) to estimate both upper and lower bounds on the divergence times of the species studied.

DNA-Hybridization Phylogeny

Marshall and Swift (accepted) provide a DNA-hybridization phylogeny (fig. 1A) of four scutelline sand dollars, *Dendraster excentricus* (Eschscholtz), *Echinarchninius*

A) DNA hybridization

B) Morphology (Mooi 1987, 1990)

FIG. 1.—Conflicting phylogenies of clypeasteroids. A), DNA-hybridization phylogeny, with percentage support in bootstrap analysis of $\Delta T_m$ values. (Marshall and Swift, accepted). B), Phylogeny derived from morphological studies of Seilacher (1979), Jensen (1981), Smith (1984), and Mooi (1987). Number of synapomorphies that support each node in Mooi's (1987, 1990) analyses are shown. The node drawn into question by the DNA hybridization data (●) is the least supported of all nodes.
parma (Lamarck), and the two mellitids Leodia sexiesperforata (Leske) and Mellita tenuis Harold and Telford. A sea biscuit, Clypeaster rosaceus (Linne), was used as an outgroup. Table 1 shows the averaged $\Delta T_m$ and $1/(\text{normalized } \% \text{ hybridization (NPH)\})$ (used as an approximation for $\Delta T_m R$; see Marshall and Swift, accepted) values.

Reliability of the DNA-Hybridization Phylogeny

Three data sets [$\Delta T_m$ (both corrected for multiple hits and not corrected for multiple hits) and $1/(\text{NPH})$] were analyzed both with and without averaging reciprocal distances, both with and without the assumption of a molecular clock, and with the algorithms of Cavalli-Sforza and Edwards (1967) and Fitch and Margoliash (1967). The topology shown in figure 1A was obtained in all these analyses, as well as when the data were jackknifed (Marshall and Swift, accepted); it would appear that the phylogenetic signal is unambiguous. As a further test of the robustness of the phylogeny, a bootstrapping technique [of Krajewski and Dickerman (1990), as modified by Marshall (1991)] was applied. The tree found in the analyses described above (fig. 1A) was obtained in 94% ($\Delta T_m$, corrected for back mutations) and 86% (1/NPH) of the 500 pseudoreplicates performed for each set of data. The conflicting morphological tree (fig. 1B) was not found among any of the pseudoreplicate trees for the 1/NPH data set and was found in just 1% of the $\Delta T_m$ pseudoreplicate trees (Marshall and Swift, accepted). The nonparametric Mann-Whitney $U$ test does not provide a reliable indication of the robustness of phylogenies derived from distance data (Marshall 1991) and was not used. Before the DNA data are considered further, two common misconceptions regarding the mode of analysis and nature of DNA hybridization data must be considered.

The Molecular Clock

It was initially thought that DNA-hybridization data had to exhibit clocklike behavior before they could be used to estimate phylogenies (e.g., see Sibley and Ahlquist 1981; Brownell 1983; Houde 1987a, 1987b; Marshall 1988). However, the assumption of a molecular clock is not required to retrieve phylogenetic information from distance data (see, especially, Springer and Krajewski 1989a; also see Sheldon 1987), and, in unusual cases, computing trees under the assumption of a molecular clock may actually lead to incorrect topologies (fig. 2).

The ability of distance algorithms to retrieve correct topologies is largely dependent

### Table 1

<table>
<thead>
<tr>
<th></th>
<th>Dendraster excentricus</th>
<th>Echinorachnius parma</th>
<th>Leodia sexiesperforata</th>
<th>Mellita tenuis</th>
<th>Clypeaster rosaceus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>D. excentricus</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>23.0 ± 2.1</td>
</tr>
<tr>
<td><strong>E. parma</strong></td>
<td>0.029 ± 0.003</td>
<td>20.7 ± 1.0</td>
<td>20.1 ± 1.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>L. sexiesperforata</strong></td>
<td>0.072 ± 0.010</td>
<td>0.059 ± 0.004</td>
<td>16.2 ± 0.9</td>
<td>21.5 ± 3.0</td>
<td></td>
</tr>
<tr>
<td><strong>M. tenuis</strong></td>
<td>0.061 ± 0.050</td>
<td>0.046 ± 0.002</td>
<td>0.022 ± 0.001</td>
<td>23.5 ± 1.2</td>
<td></td>
</tr>
<tr>
<td><strong>C. rosaceus</strong></td>
<td>0.150 ± 0.030</td>
<td>0.147 ± 0.006</td>
<td>0.166 ± 0.008</td>
<td>0.151 ± 0.006</td>
<td></td>
</tr>
</tbody>
</table>

**Note**—Mean ± standard deviations, based on four determinations, are given for $\Delta T_m$ (in °C; above diagonal) and for $1/(\text{normalized } \% \text{ hybridization})$ values (below diagonal). The $\Delta T_m$ values have been corrected for multiple hits by using the formula of Jukes and Cantor (1969) and assuming that 1°C = 1% mismatch.

**Source**—Marshall and Swift (accepted).
FIG. 2.—Data showing that assumption of molecular clock is not required to recover correct topologies by using distance data. The data matrix was derived from the “correct” topology shown in i). i), Analysis with algorithm that does not assume molecular clock (FITCH program in PHYLIP 3.2; Felsenstein 1989) but recovers “correct” topology from data matrix. ii), Analysis under assumption that molecular clock (KITSCH in PHYLIP 3.2; branch tips are contemporaneous) does not recover correct topology from data matrix. In this hypothetical example, taxon A has been designated as the outgroup.

on the distances being additive (Bledsoe and Sheldon 1989; Springer and Krajewski 1989a). Nonadditivity may result from measurement error and homoplasy (Springer and Krajewski 1989b; Bledsoe and Sheldon 1990), and while nonadditivity is more likely to result when rates of DNA sequence divergence are highly variable, rate variability per se does not negate the value of distance data.

Sympleiomorphies and DNA-Hybridization Distances

DNA-hybridization distances are a measure of the degree of mismatch between homologous strands of DNA. The word “distances” may raise the specter of “phenetics” for some, and it might be concluded that the phylogenetic value of DNA-hybridization data is compromised because of the inclusion of shared primitive characters (sympleiomorphies) in computed distances. However, only substitutions that occurred after species have become separate contribute to interspecies distances measured by DNA-hybridization techniques. Sympleiomorphies do not result in mismatches and therefore do not contribute to interspecies distances (Sibley et al. 1987; Sarich et al. 1989; Bledsoe and Sheldon 1990).

Strengths and Weaknesses of DNA Hybridization

DNA hybridization is thought to be a reliable source of phylogenetic information, in part because of the enormous amount of information assayed. A $\Delta T_m$ of 1°C between two species is roughly equivalent to a 1% difference in their single-copy DNAs (Werman et al. 1990), or about 5,000,000 bp differences for sand dollars. Thus each internode distance in the DNA-hybridization phylogeny corresponds to a sequence divergence of $\sim 10^8$ point mutations. Even if single genomic “events” cause the simultaneous appearance of thousands of point mutations, a $\Delta T_m$ of even a few degrees (as is the case in the sand dollar phylogeny) would still correspond to differences in base sequence that were the result of thousands or tens of thousands of such events.

Further support for DNA hybridization comes from the empirical observation that DNA-hybridization phylogenies usually correspond well with other phylogenetic analyses (e.g., for birds, see Krajewski (1989) and Diamond (1990) for a perspective
on Sibley and Ahlquist's work; for hominoids, see Sibley and Ahlquist (1984) and Caccone and Powell (1989); for cave crickets, see Caccone and Powell (1987); for bears, see O'Brien et al. (1985); and for marsupials, see Springer et al. (1990). Differences usually occur only where there has already been dispute in the morphological literature (as is true for *Echinarchnium* and *Dendraster*). In birds, where most DNA-hybridization data have been collected, not only do DNA-hybridization phylogenies that differ from traditional phylogenies usually find some morphological support, but, in some cases, they help clarify previously misunderstood biogeographic problems (Diamond 1983).

However, a frustration with DNA-hybridization data is that it is effectively impossible to determine the exact nature of the mismatches that contribute to the measured interspecies distances. Thus it is difficult to critically examine the data to determine whether any systematic biases or unanticipated properties of the DNA renaturation kinetics exist to warrant rejection of the data as a source of reliable phylogenetic information. This does not invalidate the use of DNA-hybridization data, but it does suggest caution in interpreting the results gained from such experiments.

In addition, our inability to determine the exact nature of the mismatches that lead to DNA-hybridization distances means that DNA-hybridization data cannot easily be combined with data derived from other sources, should such a combined analysis be desired. However, this may not be a serious disadvantage, since combining molecular and morphological data sets in a single analysis seems limited for two reasons; first, as molecular data sets become increasingly large, they are likely to numerically swamp the contribution of morphological characters. Second, and more important, combining molecular and morphological data sets in a single analysis negates the major advantage of utilizing independent data sets in evolutionary studies; by keeping the data separate, hypotheses concerning the nature of morphological evolution can be carried out without the circularity that accompanies testing hypotheses concerning morphological evolution with morphologically derived phylogenies (Olmstead 1989), and vice versa. Given both the robustness of the DNA-hybridization phylogeny vis-à-vis alternative methods of analysis and the expectation, given our current knowledge of the nature of single-copy DNA evolution, that DNA-hybridization data are phylogenetically informative, the DNA phylogeny of sand dollars (fig. 1A) deserves serious consideration.

The Morphological Phylogenies

Most morphological clypeasteroid phylogenies (Seilacher 1979; Jensen 1981; Smith 1984, p. 191; Mooi 1987) differ from the phylogeny derived from the DNA-hybridization data (Marshall and Swift, accepted); in these morphological phylogenies, *Echinarchnium* is the sister group of the other three sand dollars (fig. 1B), while in the DNA-hybridization phylogeny it is the sister group of *Dendraster* alone (fig. 1A). Of the morphological studies, only Durham's (1955, 1966) analysis gives results that are the same as those given by the molecular phylogeny. However, no explicit criteria were given by Durham (1955, 1966) in support of his evolutionary tree, and it is difficult to evaluate the strength of his scheme in light of the other analyses, and his phylogeny will be considered no further. Note that while Seilacher (1979), Jensen (1981), Smith (1984, p. 191), and Mooi (1987) all agree over the branching pattern of the four genera depicted in figure 1B, they differ from each other, often considerably, with respect to the placement of other clypeasteroid taxa.

The most complete morphological analysis of the clypeasteroids is the thorough
study by Mooi (1987), and the discussion of the quality of the morphological data is primarily based on his cladistic analysis. The relevant characters used by Seilacher (1979) and Jensen (1981) are discussed subsequently. A. B. Smith's (personal communication) characters are encompassed by those of Seilacher (1979) and Mooi (1987).

Quantity of Characters Contradicted

Figure 1B shows the number of synapomorphies that support each node in Mooi's (1987, 1990) phylogenetic analysis of the clypeasteroids relevant to the present study. If the DNA-hybridization phylogeny is correct, then it is the characters that support the node indicated by the dot (●) that must have evolved homoplasiously. This is the least-supported node in figure 1B, with just four characters, compared with the 8, 19, and 14 characters that support the other nodes. Acceptance of the DNA-hybridization data does not require any modification in our interpretation of the bulk of the morphological characters. In the currency of numbers of characters, the node drawn into question by the DNA data is exactly the node where the morphological data are weakest.

Quality of Characters Contradicted

Despite the way characters appear in data matrices, not all characters are equally informative; that is, some characters, by virtue of either their relative complexity or their limited intraspecific variability, are more likely to represent true homologues than are others. The notion of character quality is difficult to quantify, yet it is relatively easy to demonstrate. Below I examine two highly informative characters. The first supports the monophyly of the four sand dollars, and the second supports the monophyly of Mellita and Leodia.

1. Lantern supports. In the outgroup Clypeaster (and all other Clypeasterines) the lantern supports are paired structures that attach to the ambulacral basicoronal plates (fig. 3Ai), while in the four scutellines (and in laganines and all other scutellines) the lantern supports are single elements that attach to the interambulacral basicoronal (fig. 3Aii). This is a relatively complex character, involving not only the structure of the lantern supports themselves but also their position with respect to the underlying plates that make up the test.

2. Architecture of lunules. Lunules have evolved several times in sand dollars, in the rotulids, and in scutasterids, perhaps twice in the astriclypeids, and twice in the mellitids and nearest outgroups (Seilacher 1979). Figure 3B shows the position and architecture of the lunules in an astriclypeid and Mellita. Both have lunules in the posterior ambulacral, but they are constructed differently in each species. Note that if one was unaware of the difference in the construction of these lunules, the presence of the lunules could be scored (mistakenly) as a synapomorphy between the two groups.

Unlike the characters described above, the characters drawn into question by the DNA phylogeny generally have less architectural complexity and little or no specific positional relations with respect to other morphological features and are thus, on a priori grounds, more likely to be unidentified homoplasies than characters such as those described above. If the DNA-hybridization phylogeny is correct, then these characters are homoplasies and must have been either gained in parallel in Dendraster and the mellitids or gained once in a common ancestor of Dendraster, Echinarchnium,
FIG. 3.—Morphological characters with greater (A and B) or lesser (C–E) phylogenetic reliability. A, Lantern supports (heavy outline) viewed from inside test: (i) outgroup condition (*Fellaster zelandiae*) and (ii) ingroup condition (*Echinarchnus parma* is shown). The peristome (mouth) is denoted by the central blackened circle, with anterior at the top. B, Lunules: (i) crosslinked lunule in the astriclypeid *Echinodiscus bispereforatus* and (ii) festooned lunule in *Mellita quinquiesperforata* (shown) and *Leodia*. Complete specimens are viewed aborally. C, Miliary spines terminate in (i) epithelium (*Clypeaster* and *Echinarchnus*) and (ii) fluid-filled sacs (*Dendraster, Mellita*, and *Leodia*). D, Geniculate spines (arrow): (i) locomotory (left) and geniculate (right) spines from *D. excentricus* and (ii) locomotory (left) and ambulacral (right) spines from *Scaphechinus mirabilis*. E, Food grooves. Shown are (i) *Clypeaster subdepressus* (outgroup condition), (ii) *Echinarchnus parma*, and (iii) general pattern in *D. excentricus* (figured), *Mellita*, and *Leodia*. Only major branches of the food-groove system are shown in ii and iii. bp = Ambulacral basicoronal plate; lu = lunule; sc = fluid-filled sac; ep = epidermis; sk = internal calcite skeleton; mu = musculature at base of spine; t = test surface; fg = food groove; and s = sutures. Interambulacral plates are stippled. Scale bars are as follows: A and B = 10 mm; C = 400 μm; D = 1 mm; Ei and Eiii = 50 mm; and Eii = 10 mm. A–D are modified from Mooi (1989), and E is adapted from Mooi (1987).

*Leodia*, and *Mellita* and subsequently lost in *Echinarchnus*. The first four characters discussed are based on the original descriptions by Mooi (1987).

1. Sacs on miliary spine tips. In *Clypeaster* the tips of the miliary spines are covered with a thin layer of epithelium. In *Echinarchnus* (fig. 3Ci) there is a thickening of this epithelial layer. In *Dendraster* (fig. 3Cii) and the melliids the basement membrane of the epithelium has become separated from the spine tip and has become inflated with fluid. This character is simple compared with the two characters described
above; there are no subtleties of topological position and only a little detailed architecture (associated with the distribution of collagen fibers that support the sac). Moreover, fluid-filled sacs are occasionally seen on the regular sea urchin *Lytechinus* (G. A. Wray, personal communication). The reliability of a character is decreased when the intraspecific variability begins to approach interspecies variability, and thus, even though *Lytechinus* is phylogenetically far removed from the sand dollars, the occasional presence of fluid-filled sacs in *Lytechinus* adds some strength to the possibility of independent acquisition of the character within the sand dollars. Secondary loss is also a possibility, which of course would leave very little trace (or is the thickened epithelial layer the remnant of the sac?). Other than its simple presence, there is little information available to make a strong case for a single origin of the character.

2. Lengths of barrel-tipped podia. In *Clypeaster* and *Echinarchnus* the barrel-tipped podia are present in one length. In *Dendraster* and the mellitids there are both long and short barrel-tipped podia. As is the previous character, this is a simple character; there is neither positional information nor architectural detail that could be used to detect parallel or convergent evolution in this character, had it actually arisen homoplasiously.

3. Geniculate spines. In *Dendraster* and the mellitids geniculate (bent) spines are found in fields associated with the food grooves, while in the outgroup *Clypeaster* and in *Echinarchnus* the spines are not bent (fig. 3D). Not only is this a simple character, like the two previous characters, but, as can be seen in figure 3D, there is not a lot to distinguish the geniculate from nongeniculate spines (*Scaphechinus mirabilis* is shown, to emphasize that the nongeniculate spines also have a degree of bending). Homoplasy would be very difficult to detect in this character. The positional information (geniculates are only found in association with food grooves) adds some strength to the character, but the possibility that this association is a functional requirement offsets, to some extent, the importance of the positional information.

4. Differentiation in spine fields. In *Dendraster* and the mellitids the spines of the oral interambulacrum are longer than those of the ambulacra and constitute a distinct locomotory field. In the outgroup *Clypeaster* and in *Echinarchnus* there is no significant differentiation in the oral spines. As is the previous character, this, too, is a simple character. It is interesting that differentiation of the oral spines has arisen at least twice within the outgroup genus *Clypeaster*, in *C. lamprus* from the Caribbean (T. S. Hopkins, personal communication), and independently in a Japanese species (personal observation). Thus there is some evidence that this character is especially prone to homoplasic evolution.

Two other characters have been proposed that support the monophyly of *Dendraster* and the mellitids with respect to *Echinarchnus*.

5. Food grooves. Seilacher (1979) placed *Echinarchnus* as the sister group to all other scutellines, rather than as the sole sister group to *Dendraster* (Durham 1955, 1966), on the basis of the relationships between the food grooves and underlying plates. In *Echinarchnus* (and outgroups) the primary trunk of the food groove runs suturally and extends well beyond the basicoronals (fig. 3Ei and ii). In all other scutelline sand dollars the food grooves bifurcate proximally (at the ends of the basicoronal plates) and then run in the middle of the plates, rather than suturally (fig. 3Eiii). However, *Echinarchnus* shows an unusual trifurcation of the food groove distally, and at this point the food grooves leave the sutures (fig. 3Eii). The phylogenetic value of this character in *Echinarchnus* is unclear; near the mouth the food groove shows the outgroup condition of sutural growth, but distally the food groove has the derived
condition of growth across plate sutures. Depending on which aspect of the food grooves is considered, the form of the grooves in *Echinarchnus* may be viewed as either a modification of the primitive or the advanced condition! Seilacher (1979) views the *Echinarchnus* condition as a modified sympleiomorphy, while Mooi (1987) suggests that it is a modification of the derived bifurcate condition. Note the difficulty in deciding whether the condition in *Echinarchnus* is derived from the *Clypeaster* or *Dendraster* condition or whether it is intermediate between the two. Difficulty in interpreting this character renders it of questionable phylogenetic value.

6. Jensen (1979, 1981) has provided a detailed scanning-electron microscopic analysis of the tooth structure of a wide range of echinoids. She noted that *Mellita* and *Dendraster* share a unique tooth-plate construction, while *Echinarchnus* shares a different state with some other clypeasteroids (Jensen 1981). The fine structure of the plates is quite complex, suggesting a relatively strong phylogenetic signal, though understanding the evolution of the character is made difficult by the large number of described character states. There are at least six types of tooth plates known in the clypeasteroids, and there are four types of needles-prisms systems. This represents the only relatively informative character drawn into question by the DNA-hybridization data.

**Correlated Characters?**

Characters 2-5 discussed above are all directly associated with the food grooves; the geniculate spines are only found in food groove-associated spine fields, as are the barrel-tipped podia. These four characters may well be part of the differentiation between the food-groove spine fields and the locomotory spine fields and thus, for phylogenetic purposes, perhaps should be more properly treated as a single character. Under this scheme, only two characters in Mooi's analysis (fig. 1B) are drawn into question by the DNA-hybridization data! The sixth character is also associated with the feeding apparatus (the nature of the tooth plates) and thus may also be correlated with the food groove-associated characters.

**Independent Support for DNA-Hybridization Tree**

**Morphology**

From a morphological standpoint a major difficulty in accepting the DNA-hybridization phylogeny is the dearth of morphological features described in the systematic literature that support a clade of *Dendraster* plus *Echinarchnus*, to the exclusion of the mellitids. In fact, at first glance there seem to be several morphological differences between these two genera, even beyond those addressed above. However, Beadle (1989, 1990) presents a scenario where, by a simple heterochrony in test growth, the symmetrical echinarachniid-type test shape could be transformed into the asymmetrical dendrasterid-type test shape, or vice versa. Significantly, Beadle (1990) showed that, if a dendrasterid/echinarachniid transformation did occur by the proposed heterochrony, then the additional differences between the two genera alluded to above are exactly the differences one would expect to see.

Thus, while morphological characters, per se, are difficult to find that support the relationship of *Dendraster* and *Echinarchnus* [though the position of the periproct with respect to the ambulacral plates noted by Beadle (1990) is a good candidate], it is easy to see how one could be derived from the other, and this transformation is consistent with the differences in the architecture of the test observed between the two species. Further, if a major heterochronic event is responsible for the transformation
of one into the other, then it is entirely plausible that a whole suite of characters could either have been lost or have reappeared, as must have occurred if *Dendraster* and *Echinarchnitus* are sister taxa with respect to the mellitids. Further analysis of all the relevant species (many of which are not discussed either here or by Beadle), with Beadle's heterochronic mechanism kept in mind, may well produce morphological support for the topology derived from the DNA data.

Beadle's (1989, 1990) arguments that *Dendraster* is the result of a heterochronic transformation of an echinarachniid are strengthened by an examination of the fossil record (Beadle 1990). There are numerous extinct taxa known from the northern Pacific that have been variously assigned to the Dendrasteridae, Echinarchnidae, or, under some phylogenetic schemes, to a third, closely allied family (Durham 1955; Nisiyama 1968; Wagner 1974; Mooi 1987). Part of the difficulty in reaching a consensus on how to classify the complex is the presence of several chimeric forms among the fossil taxa, which suggests a close relationship between the Echinarchnidae and the Dendrasteridae. A thorough revision of the northern Pacific taxa, as well as a reanalysis of the mellitids and their fossil sister groups in light of Beadle's suggestion of heterochronic transformations between members of the Echinarchnidae and the Dendrasteridae, is required to secure Beadle's morphological arguments in favor of *Echinarchnitus* being more closely related to *Dendraster* than to either *Mellita* or *Leodia*.

Paleobiogeography

For the first half of their respective histories, the dendrasterid/echinarachniid clade and the lineage that led to the mellitids occupied widely separated geographic realms. In the Eocene the first echinarachniids (*Kewia*) are known from Oregon (Linder et al. 1988), and a now-extinct sister group to the mellitids, *Monophoraster* (see Mooi 1987), appears in southern South America (Larrain 1984). In the Oligocene, a period of poor fossilization for echinoids (Raup 1975), the echinarachniid *Kewia* is known from the northwestern United States, Sakhalin, and Japan (Nisiyama 1968; Linder et al. 1988), while *Monophoraster* is still restricted to South America (Patagonia; W. J. Zinsmeister, personal communication). In the Miocene the distributions begin to converge as the mellitids become centered in the Caribbean, and since the Pliocene there has been limited overlap between some mellitids (*Encope* and *Mellita*) and *Dendraster* in southern California and Baja California (Mooi 1989; Beadle, accepted).

Thus, if *Dendraster* is really the sister group to the mellitids, then dispersal of some primitive dendrasterids from the northwest Americas to southeast South America (during a time well before dendrasterids first appear in the fossil record!) needs to be hypothesized. While not impossible (the Panamanian Isthmus was open, and modern representatives of these groups all have feeding planktrophic larvae), the spatially and temporally disjunct origins of the dendrasterids and mellitids present a major problem for the morphological cladogram—but not for the DNA-hybridization phylogeny or for Durham (1955, 1966) and Beadle (1989, 1990). For an explanation of the distribution of these sand dollars, it is more parsimonious to accept a vicariance model rather than a dispersal model.

Conclusions

In the present study I have compared the reliabilities of DNA-hybridization and morphological phylogenies of some clypeasteroids and have found that, where the two trees differ, the DNA phylogeny is more reliable. In other studies it may be found that morphological data are more reliable than are DNA data. For example, an 18S rRNA
phylogeny of amniote classes (Hedges et al. 1990) was found to differ from a robust paleontological phylogeny of the same taxa. An examination of the DNA sequences led to the identification of substantial substitution biases in the 18S rRNA sequences, and reanalysis of the sequences by the weighted parsimony algorithm of Williams and Fitch (1990) greatly improved the concordance between the paleontological and DNA phylogenies (Marshall 1992). In attempting to understand the nature of conflicting molecular and morphological phylogenies (rather than taking grand averages of the available information), one can gain insight into the nature of evolutionary changes in both morphological and molecular data sets.

With regard the clypeasteroids, while I have argued that it is only a few of the weakest of characters that are drawn into question by the DNA tree, it does not follow that those characters should not have been used in the morphological analyses in the first place. In the absence of any other information—i.e., the conflicting DNA and biogeographic data, as well as the morphological analyses of Beadle (1989, 1990)—these characters represent good hypotheses of relationship. This examination of the quality of the morphological characters used by Mooi (1987) does not negate the fact that contradictory evidence exists on the affinities of *Dendraster* and *Echinarchnius* and that perhaps the DNA-hybridization phylogeny is in error, though I have argued that this latter possibility is unlikely. DNA sequence data are now being collected to test the veracity of the DNA-hybridization and morphological data, and further morphological analysis is also required.

The DNA-hybridization phylogeny implies that morphological evolution among the four sand dollars has proceeded in a (slightly) less than most parsimonious way. Parsimony is the most commonly used criterion for deciding between phylogenetic hypotheses, especially between competing morphological hypotheses. Thus, to some, the fact that the DNA data support a less than most parsimonious tree for the sand dollars may indicate some error in analysis of the DNA data. But it should be noted that there is no a priori reason for favoring the most parsimonious tree simply because it is the most parsimonious (see, especially, Sober 1988, p. 104; also see Wiley 1975; Johnson 1982; Marshall 1986; Sluys 1989). To quote a philosopher: "Unless we give more details, there is no more reason to give pride of place to a principle of parsimony than to a principle of plenitude" (Sober 1988, p. 104).

The DNA-hybridization data give reason to favor a little more plenitude in interpreting the morphological data than do the morphological data alone, and, as shown above, when standard criteria for recognizing homologies are employed to assess the relative informativeness of the morphological characters implied to be homoplasies by the DNA-hybridization phylogeny, the supposition that they represent undetected homoplasies is not unreasonable. Biogeographic data and a detailed analysis of the morphologies of *Dendraster* and *Echinarchnius* in the light of heterochronic mechanisms further support the probability of a little more plenitude.

Acknowledgments

Thanks to Drs. Rich Mooi, Steven Beadle, Bruno David, Walter Fitch, two anonymous reviewers, and especially Clint Turbeville for critical comments on the manuscript. Thanks also to Drs. Steven Beadle, Rich Mooi, and Andrew Smith for providing access to pre- and unpublished material. This work was supported by a Fellowship from the Searle Foundation and by a grant from the Eli Nierman Foundation. Thanks to Rudy Raff and the Department of Biology at Indiana University for providing the
facilities for preparation of the manuscript. The manuscript was prepared under support from NIH NRSA Fellowship 5 F32 HD07332-02.

LITERATURE CITED

—. 1990. The origins of the family Dendrasteridae (Echinoidea, Clypeasteroida): rapid evolutionary change through heterochrony. Ph.D. diss., The Johns Hopkins University, Baltimore.
—. The biogeography of origin and radiation: dendrasterid sand dollars in the northeastern Pacific. Paleobiology (accepted).
DNA and Morphological Phylogenies of Sand Dollars


—. 1990b. The fossil record and estimating divergence times between lineages: maximum divergence times and the importance of reliable phylogenies. J. Mol. Evol. 30:400–408.


MARSHALL, C. R., and H. SWIFT. DNA-DNA hybridization phylogeny of sand dollars and highly reproducible homologous extent of hybridization values. J. Mol. Evol. (accepted)


WALTER M. FITCH, reviewing editor

Received January 16, 1991; revision received August 28, 1991

Accepted September 3, 1991
Rapid Surveying of DNA Sequence Variation in Natural Populations

Enrique P. Lessa
Museum of Vertebrate Zoology, University of California, Berkeley

DNA sequencing can be costly and time consuming for population studies because of the relative rarity of variation along exons. These problems can be substantially reduced by the use of the polymerase chain reaction on introns using primers from the exon region. These problems can be further reduced by the use of denaturing gradient gel electrophoresis to identify those alleles in need of sequencing.

Introduction

The study of variation of DNA sequences promises to provide important insights into the genetic structure and dynamics of natural populations. However, sequencing is too costly and time consuming for the large samples and multiple loci often required for population-level studies. In addition, heterozygotes are difficult to score in sequencing gels; this complicates the assessment of genotypic variation in nuclear loci. Here I present an approach for overcoming these two problems, capitalizing on the known intron/exon structure of many eukaryotic genes and on available laboratory techniques.

General Approach

The coding segments (exons) of most eukaryotic genes are interrupted by non-coding introns. The latter are likely to vary within species and can therefore be targeted as markers of population variation and subdivision, by using conserved regions of flanking exons to design primers for amplifications via the polymerase chain reaction (PCR; Saiki et al. 1988). The analysis of these double-stranded PCR products usually proceeds by asymmetric amplifications (or strand separation) followed by direct sequencing of both strands. However, the gathering of allelic data may be simplified by scoring individual genotypes by using denaturing gradient gel electrophoresis (DGGE) and limiting the sequencing efforts to the distinct alleles identified by this process [fig. 1(a)].

Detailed descriptions of laboratory equipment and procedures for DGGE are readily available (Myers et al. 1987, 1989a, 1989b; Sheffield et al. 1989; Abrams et al. 1990). In brief, DGGE allows double-stranded PCR products that differ by one or more mutations to be separated on acrylamide gels cast with linear gradients of denaturants (urea and formamide). This is possible because mutations alter the DNA segments’ resistance to denaturation, resulting in changes in mobility on gradient gels. As shown diagramatically in figure 1, there are two types of denaturing gradient gels:

1. Key words: polymerase chain reaction, denaturing gradient gel electrophoresis, intron, β-globin.

Address for correspondence and reprints: Enrique P. Lessa, Museum of Vertebrate Zoology, University of California, Berkeley, California 94720.

© 1992 by The University of Chicago. All rights reserved.
0737-4038/92/0902-0011$02.00
DNA extractions

30

double-stranded amplifications

60 12
asymmetric amplifications
(or strand-separation)

60 12
direct sequencing

30

DGGE

(a)

(b)
PERPENDICULAR DGGE

(c)
PARALLEL DGGE

0% 50% 80%

40%

60%

aa ab bb

FIG. 1.—(a), Flow chart depicting use of DGGE in combination with PCR and direct sequencing. DGGE represents a detour from the usual procedures that allows one to score individual genotypes and to limit sequencing efforts. In the example presented in the present paper, 30 specimens would have required 60 asymmetric amplification/sequencing reactions. DGGE showed that only three alleles were present in the samples; sequencing replicates of each required a total of 12 asymmetric amplification sequencing reactions. (b), Diagrammatic representation of perpendicular DGGE of single heterozygous specimen. The point of partial denaturation of the two alleles is shown at ~50% denaturant. (c), Same two alleles separated and scored on parallel gels with ranges of denaturant centered at ~50%.
1. Perpendicular gradient gels are used to evaluate the response of a single PCR product to the gradient. The gel has a full range (0%-80%) of denaturants from side to side, i.e., perpendicular to the direction of electrophoresis. Perpendicular DGGE is used to determine the point of partial denaturation of the DNA segment under study [fig. 1(b)].

2. Parallel gradient gels are cast with a gradient of denaturants from top to bottom, i.e., parallel to the direction of electrophoresis. The denaturants on parallel gels cover a range of concentrations of 20%-30% points centered about the point of denaturation previously determined by perpendicular DGGE. On this narrower range and under constant temperature, parallel DGGE allows the separation of PCR products differing by one or more mutations, because such mutations change the DNA’s resistance to denaturation. In a heterozygous sample, for instance, each of the alleles will partially denature (and consequently slow down its rate of migration) at a different point along the gradient. Multiple samples can be loaded and run simultaneously on parallel denaturing gradient gels and can be scored after ethidium bromide staining [fig. 1(c)].

In this context, DGGE may be conceived of as a detour from the usual PCR/sequencing procedures [fig. 1(a)], one that both simplifies the scoring of genotypes and limits costly and labor-intensive asymmetric amplifications (or strand separation) and sequencing to the minimum. Also, alleles are effectively gel purified in parallel DGGE. If the sequences of alleles only found in heterozygotes are of interest (e.g., rare alleles in hybrid zones), they can be cut out of the parallel denaturing gradient gels for subsequent reamplification and sequencing.

If no sequences are available for the loci and species to be studied, the approach outlined above for the study of introns can be best implemented in several steps, each using a specific combination of primers (fig. 2). First, a pilot PCR/sequencing project directed at obtaining a few preliminary sequences may be carried out with primers designed on the basis of conserved exon regions identified in alignments of sequences of other taxa. The preliminary sequences should include data from at least one of the exons flanking the targeted intron, in order to verify the identity of the amplified gene and to design one species-specific primer for future experiments. This primer is synthesized in two versions. One version should include a 40-mer “GC-clamp” on the 5' end, to allow maximal resolution in DGGE (Myers et al. 1989a). It is used both for double-stranded PCR amplifications of all specimens and for the screening of amplified products for allelic variation in denaturing gradient gels. Another version of the primer,

Exon 1    Intron 1    Exon 2

Primer 1

Primer 3  Primer 2

Fig. 2.—Diagrammatic representation of intron and flanking exons and of primers used for uncovering variation in intron by combination of amplifications by PCR, DGGE, and direct sequencing (see text for details). Primers 1 and 2 are based on conserved regions identified in alignments of available sequences of the target gene in other species and are used to obtain initial sequences. Primer 3 is species specific and is designed in two versions. One has a GC-clamp on the 5' end and is used in conjunction with primer 1 for double-stranded PCR amplifications of products to be screened by DGGE. A version of primer 3 without the GC-clamp is used with primer 1 to reamplify and directly sequence the alleles identified by DGGE.
one that lacks the GC-clamp, is used both for asymmetric amplifications and for direct sequencing of the amplified products.

An Example: β-Globin Intron 1 in Pocket Gophers

This approach will be illustrated with an analysis of variation in intron 1 of the adult β-globin gene of pocket gophers (*Thomomys bottae*), surveyed in 30 individuals from eight localities in the lower Colorado River area. DNA was obtained from frozen liver samples by sodium dodecyl sulfate/proteinase K/RNAs enzyme followed by phenol or sodium chloride extraction and alcohol precipitation (Maniatis et al. 1982, pp. 458–462; Miller et al. 1988).

The two initial primers were designed on the basis of alignments of known sequences of genes of the β-globin family in a variety of mammals. The 3' ends of the three exons showed greater conservation than did other regions. Intron 1 of adult β-globin is consistently small (~120 bp long) and was targeted for initial amplifications with two primers—β-1 (5'-GTTGGGTGGTGGAGGCCTGCCCAG) and β-2 (5'-CCTGAGCTCTCAGGATCCACATGCA)—complementary to segments of exons 1 and 2, respectively.

The initial sequences of four pocket gopher samples showed greatest similarity to adult β-globins of various mammals, followed by embryonic β-globins and by pseudogenes in the family. The pocket gopher sequences of exon 2 were used to design primer β-6 (5'-TTCCGCTCAGACTGGTAAAGGTGCCCTTCA). The GC-clamp attached to the 5' end of one version of β-6 (β-6+GC) had the following sequence: 5'-CGCCCGCCGCGGCCCCGGCGGCCGCCCCGCCCCGCCCCGCCC. Myers et al.'s (1989a) suggestion not to include bases other than G and C was followed, causing the GC-clamp to differ from theirs by only one position.

The PCR products amplified with β-1 and β-6+GC were used for DGGE. First, the point of partial denaturation of a single amplified product was determined to be 55% on a perpendicular denaturing gradient gel (Myers et al. 1989a). Subsequent parallel denaturing gradient gels, used to screen allelic variation in the sample, were cast with linear gradients of 45%–65% urea. All gels were run for 3–5 h at 150 V while immersed in an aquarium at 60°C and were flushed with water and stained with ethidium bromide.

Three alleles were identified by DGGE (fig. 3). Two of them were restricted to one of the two “genetic groups” (defined by allosyme analysis; Smith and Patton 1980, 1984) represented by these populations, but the third was found in both (fig. 4). The distribution of β-globin alleles is consistent with the pattern uncovered by protein electrophoresis in that (1) variation in allele frequencies may be marked across short geographic distances and (2) whereas the two genetic groups differ by substantial

![Fig. 3.—Negative image of parallel denaturing gradient gel stained with ethidium bromide and showing allelic combinations found in present study. Genotypes (from left to right) are aa, aa, ab, ab, bb, bc, bc, cc, and cc. The arrow indicates the direction of migration during electrophoresis.](image-url)
Asymmetric amplifications (Gyllensten and Erlich 1988) were carried out using alleles sampled from either agarose gels or acrylamide denaturing gradient gels. Direct sequencing of the reamplified alleles showed that intron 1 of the adult β-globin of pocket gophers is 126 bp long and that two point substitutions account for the differences between the three alleles (fig. 5). No additional variation was detected, either by sequencing additional samples or by the introduction of “heteroduplex analysis” to the DGGE protocol. In general, however, heteroduplex analysis is advisable (see Abrams et al. 1990, and references therein).

The results attest to the resolving power of DGGE. Furthermore, because the observed differences between the three alleles are simple, a direct relationship can be drawn between such differences and the behavior of the alleles on the gradient gels. Thus, increasing resistance to denaturation, from allele a to allele b and allele c, is
accompanied by replacements of AT by GC base pairs (figs. 3 and 5). Also, the data show that even small introns may be variable at the population level.

Whether the level of variation found in this gene and taxon is representative of other introns remains to be determined. In principle, it seems desirable to target larger introns; DNA segments ≤ 500 bp are within the range of maximal resolution of DGGE (Mycrs et al. 1989a, and references therein).

## Extensions, Limitations, and Alternatives

This approach is applicable to a broad spectrum of species and loci because PCR can withstand mismatches in the primers (Kwok et al. 1990); if no sequences are known for the species of interest, available sequences of other taxa will often suffice for designing primers (Kocher et al. 1989). Like allozyme electrophoresis, the combination of PCR, DGGE, and direct sequencing provides genotype-frequency data to population geneticists, conservation biologists, and other students of genetic diversity in the wild. Working with these methods at the DNA level, however, has distinct advantages. First, it provides greater sensitivity because of the ability to target a broader range of loci and to detect variation in noncoding regions and in silent mutations in exons. Second, information on sequence variation in nuclear genes allows the construction of “gene trees” and, more generally, the use of phylogenetic approaches to population genetics (Slatkin 1989; Slatkin and Maddison 1989), which have thus far been virtually limited to mitochondrial DNA data (Avise 1989). Third, the methods can be applied to DNA sources other than frozen tissues, including museum specimens (Thomas et al. 1990). Finally, the protocol can easily be extended to other cases in which large samples and/or sorting of variation within individuals are desired, such as population-level surveys of mitochondrial DNA (including the detection of heteroplasmy) and studies of sequence variation within gene families. The main limitations of this approach are (a) the costs involved in primer synthesis, PCR, and sequencing; (b) the need to target relatively small (≤500 bp) segments, in order to achieve full resolution in DGGE; and (c) the need to have sequence data for the target genes of related taxa, in order to design primers.

Obviously, choosing between this and other approaches depends on the research project. Among the possible alternatives, no method can match allozyme electrophoresis in terms of cost efficiency and the ability to study multiple loci. If working at the DNA level is preferred, alternatives include at least the following: (a) temperature gradient gel electrophoresis (Wartell et al. 1990), conceptually very similar to DGGE, but using gradients of temperature rather than denaturants; (b) single-stranded con-

---

### FIG. 5—Diagrammatic representation of intron 1 of adult α-globin of pocket gophers. The intron is 126 bases long in the three alleles, which, as noted, differ only at positions 24 and 97.

<table>
<thead>
<tr>
<th>Allele</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>b</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>c</td>
<td>G</td>
<td>C</td>
</tr>
</tbody>
</table>

---

328 Lessa
formational polymorphisms (Orita et al. 1989), primarily for short (<200 bases) PCR products, especially if full resolution is not required; and (c) restriction-fragment analysis, which can be combined with DGGE to achieve increased levels of resolution (Myers et al. 1989a, and references therein).

Sequence Availability

Sequences have been deposited in GenBank under accession number M62863 and are also available from the author.

Acknowledgments

I am grateful to James L. Patton, Kelley Thomas, Maria P. Echeverria, and Cristian Orrego for helpful discussions; to Karen Klitz for drawing the figures; and to Brian Charlesworth, Craig Moritz, Barbara R. Stein, and two anonymous reviewers for comments on the manuscript. This work was supported by NSF grant BSR-88-22345 (to J. L. Patton) and by an Alfred P. Sloan Postdoctoral Fellowship in Molecular Studies of Evolution.

LITERATURE CITED


BRIAN CHARLESWORTH, reviewing editor

Received February 26, 1991; revision received September 4, 1991

Accepted September 9, 1991
Sequence Instability and Functional Inactivation of Murine Y Chromosomes Can Occur on a Specific Genetic Background

Ferez S. Nallaseth
Department of Cell and Developmental Biology, Roche Research Center

When the Y chromosome from Mus. poschiavinus (Y^pos) is backcrossed onto the C57BL/6J laboratory strain, testicular dysfunction occurs at high frequencies. When five different multicopy probes from the recombinationally suppressed region of the Y chromosome were used, genomic DNAs from sibling female progeny of C57BL/6J Y^pos males were found to contain Y^pos-specific sequences ranging from trace levels to levels consistent with an intact Y chromosome. Females with a high copy number of Y^pos-specific sequences had a karyotype of XY^pos and were sterile. Females with trace levels of these sequences were XX and fertile. Repeated sequences in the testis-determining-region (Sxr) of inactive Y^pos chromosomes were unstable relative to sequences in non-Sxr regions. In contrast, the Y^pos chromosome was stable and functioned normally in other inbred laboratory strains such as 129/Sv. The frequency and extent of Y^pos chromosome instability increased with successive backcrosses from stable (129/Sv) to unstable (C57BL/6J) genetic backgrounds. Traces of Y^pos-specific sequences were first detected in N2 female offspring of F1 males. Therefore, sequences were deleted from Y^pos chromosomes in the F1 male germ line and were transmitted to N2 females; inactive Y^pos chromosomes (XY^pos females) were first detected in the N3 generation. The mouse line being derived by backcrossing the Y^pos chromosome onto C57BL/6J inbred strains ended in the N7 generation, since all XY^pos offspring were sterile. Even stable repeated sequences from the non-Sxr regions of their inactive Y^pos chromosomes were precisely rearranged in these N7 offspring at high frequencies. These data are consistent with hybrid dysgenesis in mammals.

Introduction

Haldane first observed that, when hybrids of evolutionarily diverged species or races are either sterile or inviable, they belong to the heterogametic (XY or ZW) sex, rather than to the homogametic (XX or ZZ) sex (Haldane 1922). It was subsequently shown that both genic and chromosomal incompatibilities were responsible for inviability and infertility in hybrids of Drosophila species (Sturtevant 1939; Dobzhansky 1951).

Furthermore, it is recognized that the heterogametic sex- and genetic-background-dependent induction of specific genotypic and phenotypic traits constitutes hybrid dysgenesis in Drosophila (Hiraizumi 1971; Kidwell et al. 1977; Bregliano et al. 1980; Bingham et al. 1982). A major genotypic trait characterizing hybrid dysgenesis is the

1. Key words: Y chromosome instability, XY females, murine hybrid dysgenesis.

Address for correspondence and reprints: Ferez S. Nallaseth, Department of Cell and Developmental Biology, Roche Institute of Molecular Biology, Roche Research Center, 340 Kingsland Avenue, Nutley, New Jersey 07110-1199.

© 1992 by The University of Chicago. All rights reserved.
0737-4038/92/0902-0012$02.00

331
generalized induction of recombination (Hiraizumi 1971). This loss of control results from the availability of breakpoints at sites of excision of transposons but also includes suppression of recombination, transmission-ratio distortion, translocations, aneuploidies, and high frequencies of mutations (Hiraizumi 1971; Kidwell et al. 1977; Hinton 1979; Bregliano et al. 1980; Bingham et al. 1982; Yannopoulos et al. 1987). Mutations of gonadal functions are also detected frequently in hybrid genomes undergoing random recombination because of (1) their nonlethal effects, (2) the large numbers of encoding target loci, and (3) the simplicity of monitoring mutant phenotypes (e.g., by sex ratios, external genitalia, fertility, and karyotypic analysis) (Kidwell et al. 1977; Bingham et al. 1982; Yannopoulos et al. 1987).

Similar genotypic and phenotypic traits have been reported in hybrids or strains derived from hybrids of several murine species. Chromosomal aneuploidy and embryonic inviability (Gropp et al. 1972; Cattanach and Moseley 1973; Capanna 1982), sterility (Gropp et al. 1972; Cattanach and Moseley 1973; Forejt and Iványi 1975; Capanna 1982; Handel 1987), and sex reversal (Eicher and Washburn 1986) occur at high frequencies. Testicular dysfunction in these hybrids/strains, however, has not been molecularly correlated to mutations resulting from induction of sequence instability. The presence of major testicular determinants on the murine Y chromosome has now been established by the canonical loss-of-function/gain-of-function analyses. This locus is referred to as testis-determining Y-linked, or Tdy (Eicher and Washburn 1986), and is contained within the Sxr region of the murine Y chromosome, which on translocation causes XX maleness (Cattanach et al. 1982; Evans and Burtenshaw 1982; McLaren and Monk 1982; Singh and Jones 1982; McLaren et al. 1988 and Roberts et al. 1988). The necessity for sequestering the Y-linked copies of the testis-determining locus from the female genome is illustrated by the dominant maleness of the sterile translocation mutant XX<sup>Sxr</sup> (Cattanach et al. 1982; McLaren and Monk 1982; Singh and Jones 1982). This sequestration is thought to be effected by the male-limited haploidy and by the suppression of interchromosomal X-Y recombination over the predominant length of the Y chromosome (Tres 1977; Chandley et al. 1984).

High frequencies of sex reversal (XY females) result when the Y chromosome from the feral species <i>Mus poschiavinus</i> (Y<sub>Pos</sub> chromosome) is backcrossed onto the C57BL/6J genetic background (Eicher et al. 1982). When this same Y<sub>Pos</sub> chromosome is backcrossed onto the DBA/2J or 129/Sv genetic backgrounds, testicular function is unaffected and normal ratios of male and female progeny are obtained (Eicher et al. 1982; Nallaseth 1987). Testicular dysfunction may result from improper reception by (C57BL/6J) autosomal loci of trans-activating testis-determining signals from the <i>Tdy</i>-locus of Y<sub>Pos</sub> chromosomes (Eicher et al. 1982). However, some of the molecular analyses of genomic DNAs from female offspring of C57BL/6J Y<sub>Pos</sub> males provide an alternative explanation. It is shown here that random mutations due to a generalized loss of control of recombination apparently result in the observed testicular dysfunction in C57BL/6J Y<sub>Pos</sub> mice. Most genotypic and phenotypic traits characteristic of hybrid dysgenesis in other species (Kidwell et al. 1977; Hinton 1979; Bregliano et al. 1980; Bingham et al. 1982; Yannopoulos et al. 1987) are also observed in (<i>M. musculus</i> × <i>M. poschiavinus</i>) F1 hybrids and in various lines derived from them (Gropp et al. 1972; Cattanach and Moseley 1973; Capanna 1982; Eicher et al. 1982; Magnuson et al. 1985; Nallaseth 1987). These data collectively argue that the C57BL/6J Y<sub>Pos</sub> mouse represents the first case of mammalian hybrid dysgenesis recognized at the level of molecular resolution.
Material and Methods

Mice

$XX^{Sv}$ males (C57BL/6J genetic backgrounds) and pure strain C57BL/6J and
129/Sv were obtained from Jackson Laboratories. 129/Sv $Y^{Pos}$ and C57BL/6J $Y^{Pos}$
males were initially transferred from the colonies of J. Barry Whitney III to the colonies
at the University of South Carolina. Mus poschiavinus (Valende) and M. domesticus
(Centreville Light) were provided by Michael J. Potter.

Southern and Densitometric Analysis

Genomic DNAs were extracted from livers by using a procedure modified from
published methods (Nallaseth and Dewey 1986; Nallaseth 1987). DNAs were digested
as per instructions, with restriction endonucleases that are commercially available.
Digested DNAs were heat inactivated in 10 mM ethylene diaminetetraacetate (EDTA)
and were ethanol precipitated, electrophoresed through (0.8%) agarose, 40 mM Tris,
20 mM sodium acetate, 1 mM EDTA pH 8.0, and 1 $\mu$g ethidium bromide/ml and
were transferred either to nylon (N-Bond; Amersham) or, occasionally, to nitrocellulose
(BA45; Schleicher and Schuell) filters. Hybridization and washing protocols are as
described elsewhere (Nallaseth and Dewey 1986; Nallaseth 1987). Nomenclature of
genomic clones and of the various Y-specific fragments subcloned from them are as
described elsewhere (Nallaseth and Dewey 1986; Nallaseth 1987). EcoRI inserts of
the Y-specific plasmid subclones were labeled either by replacement synthesis with
T4 DNA polymerase or by nick-translation. (GATA)$_5$ oligonucleotides were end-
labeled with T4 polynucleotide kinase. Some single-stranded probes of Y-specific se-
quences cloned into M13 were probes of choice. Details of choices of labeling protocols
are as described elsewhere (Nallaseth 1987). Multiple exposures of autoradiographs
were necessary for identifying and quantifying sequences that either (a) were 10-fold
different in copy numbers or (b) because of nucleotide sequence mismatches annealed
poorly with the probe. High- and low-level exposures of autoradiographs were attained
by varying the following conditions: times and temperatures (-70°C or ambient) of
exposure, use of intensifier screens, and preflashed X-ray films. XAR-5 (Kodak) X-
ray films were developed in X-O-MAT (Kodak) processors. Restriction fragments of
known and comparable copy numbers were routinely included as internal standards
in densitometric quantifications. Densitometric analysis of X-ray films was performed
in a Helena Laboratories densitometer, and the signal was analyzed with an Apple
computer and Chromatochart programs (Nallaseth 1987).

Results

Locations of Y-linked Sequences Used as Probes, and Their Advantages
for Detecting Aberrant Recombination Products

Because a generalized loss of genetic control of recombination affects multiple
loci, noncoding repeated sequences are the probes of choice for identifying the resultant
genotypic effects. An important and frequent technical complication of repeated se-
quence probes is the inability to detect minor aberrant recombination products in the
presence of a vast excess of normal copies. Y-linked repeated sequences are sensitive
probes because, owing to their male-limited haploidy and their location in the recom-
binationally suppressed chromosomal region, they allow the circumvention of this
complication. Minor aberrant interchromosomal recombination products from the
Y chromosome can be unambiguously detected in genomic DNAs of XX females.
Several Y-specific EcoRI fragments are present at 90–170 copies/male genome (Nallaseth et al. 1983; Nallaseth and Dewey 1986; Nallaseth 1987). Sequences hybridizing to cloned copies of these fragments used as probes collectively represent a minimum of 15% of the murine Y chromosome (fig. 1A). This was determined by the product of the sizes of the inserts of genomic (lambda) clones and the copy numbers of the Y-specific probes subcloned from them. The sizes and the copy numbers of the major restriction fragments hybridizing to each of these Y-specific probes are listed in figure 1A. Each of these Y-specific EcoRI fragments cross-hybridized with 1–5 copies/haploid genome of autosomal or X-chromosomal EcoRI fragments (Nallaseth and Dewey 1986; Nallaseth 1987).

Three of these Y-specific probes—A, B, and C—failed to detect differences in restriction-endonuclease fragments between Y<sub>pos</sub> chromosomes in feral Mus poschiavinus (Valende) males and 129/Sv Y<sub>pos</sub> (backcrossed: N 18) males (Nallaseth 1987). This confirmed that the Y chromosomes in laboratory strains 129/Sv Y<sub>pos</sub> and C57BL/6J Y<sub>pos</sub> mice were derived from M. poschiavinus (Valende). These probes identified three distinct Y chromosomes in M. poschiavinus, M. domesticus (Centreville Light), and M. musculus (C57BL/6J) (Nallaseth 1987).

The Sxr region and the pseudoautosomal region are located, respectively, on the short arm near the centromere and on the long arm in the telomere of the murine Y chromosome (McLaren et al. 1988; Roberts et al. 1988) (fig. 1A). These two regions permit the cytogenetic localization of our probes. EcoRI-digested genomic DNAs from XX<sub>Sxr</sub> males were hybridized with probes A–C. They failed to detect Y-specific EcoRI fragments (data not shown). The expected cross-hybridizing X-linked fragments of probe A (Nallaseth and Dewey 1986; Nallaseth 1987) were clearly detected. These three Y-specific sequences were absent from the Sxr region. A fourth probe, (GATA)<sub>5</sub>, identifies HaeIII or AluI fragments of ~23 kb and ≥100 kb that are concentrated in the Sxr region and absent in XX females (Singh et al. 1984). Therefore, I monitored the structural integrity of the repetitive sequence fraction of the Sxr region by using the (GATA)<sub>5</sub> oligonucleotide (fig. 1A).

The frequency of X-Y exchange in the pseudoautosomal region is reflected as the frequency of females inheriting Y-chromosomal restriction fragments (Cooke and Smith 1986; Goodfellow et al. 1986; Harbers et al. 1986; Rouyer et al. 1986; Simmler et al. 1987). Y-chromosomal sequences hybridizing to probes A–C were tested for their frequency of transfer to X chromosomes of inbred strains. Major Y-specific EcoRI fragments of expected size (0.6 kb) and copy number (~100) were present (++) in male genomic DNA (fig. 1B, lane 11) but absent (−) in all female genomic DNAs (fig. 1B, lanes 1–10) hybridized with probe A. Probe A also hybridized to several minor Y-specific EcoRI fragments (fig. 1B, lane 11). Although the background at 0.6 kb in one female genomic DNA (fig. 1B, in lane 10) was high, independent hybridizations have confirmed the absence of Y-specific EcoRI fragments in this DNA. Cross-hybridizing EcoRI fragments of 2.2 kb, 1.7 kb, 1.4 kb, and 0.8 kb were present in all female DNAs (fig. 1B, lanes 1–10). These fragments served as internal standards for probe A. After densitometric quantitation and correction for sizes of probes, the copy numbers of the 2.2-kb and 1.7-kb EcoRI fragments of probe A were estimated...
to equal the unique-copy levels of the 12-kb EcoRI fragment of α-globin (fig. 1C, lanes 1 and 2). The other cross-hybridizing fragments of probe A were present at lower copy numbers and were detected at higher exposures.

Major Y-specific EcoRI fragments of expected sizes (4.3 kb and 3.7 kb) and copy numbers (~100) were present (+++ ) in male genomic DNAs (fig. 1D, lanes 12 and 13) but absent (−) in female genomic DNAs (fig. 1D, lanes 1–11) hybridized with probe B. Probe B also hybridized to minor Y-specific EcoRI fragments of 2.2 kb and 6.6 kb (fig. 1D, lane 13). At high exposures, a unique-copy 3.9-kb EcoRI fragment cross-hybridizing with probe B was detected in all female DNAs (fig. 1D, lanes 1–11). This fragment served as the internal standard for probe B.

Major Y-specific EcoRI fragments of expected size (6.6 kb) and copy number (~100) were present (+++ ) in feral male genomic DNA (fig. 1E, lane 10) but absent (−) in all female genomic DNAs (fig. 1E, lanes 1–9) hybridized with probe C. Probe C also hybridized to several minor Y-specific EcoRI fragments (fig. 1E, lane 10). The multicopy 1.8-kb EcoRI fragment of probe C was not detected, as it is an RFLP (restriction-fragment-length polymorphism) that is specific for the inbred-strain (C57BL/6J) Y chromosome. At high exposures, a unique-copy 3.5-kb EcoRI fragment cross-hybridizing with probe C was detected in all female genomic DNAs (fig. 1E, lanes 1–9). This fragment served as the internal standard for probe C.

An additional 36 sibling females also failed to reveal Y-specific fragments hybridizing to probes A–C (Nallaseth 1987). Thus, if, because of pseudoautosomal exchange, Y-specific sequences hybridizing to these three probes normally undergo transfer to karyotypically XX females, they do so at frequencies of <5% (P < 0.09). In addition, the X-linked fragments detected by probe A were monomorphic among C57BL/6J, in six other laboratory strains, and in M. poschiavinus and M. domesticus (Nallaseth 1987). The likelihood that these X-linked sequences were pseudoautosomal was further reduced when five additional restriction endonucleases failed to generate fragments characteristic of Y-chromosomal sequences (Nallaseth 1987). At least 200 independent females failed to show differences in these X-linked sequences, illustrating their extreme stability.

Considerable (>90%) sequence identity existed between these Y-linked fragments and published Y-specific sequences known to be distributed along the length of the Y chromosome (author’s unpublished data). These results showed that the most likely location of the Y-linked repeated sequences hybridizing to probes A–C was in the entire, recombinationally suppressed Y-chromosomal region between the Sxr region and the pseudoautosomal region (fig. 1A). Therefore, the suitability of probes A–C for the experiments presented below was well established.

Variable Amounts of Y Pos-specific Sequences in Female Sibling Progeny of C57BL/6J Y Pos Males

Previous karyotypic analysis had failed to identify gross structural alterations of the Y Pos chromosome in C57BL/6J Y Pos mice (Eicher et al. 1982). However, others had reported that control of segregation and recombination of autosomes and X chromosomes was disrupted in mice derived from M. musculus × M. poschiavinus hybrids (Gropp et al. 1972; Cattanach and Moseley 1973; Capanna 1982; Magnuson et al. 1985). The possibility of structural instability of the Y Pos chromosome in C57BL/6J Y Pos mice was, therefore, pursued with the more sensitive approach of Southern analysis with Y-linked probes. As predicted by the instability of the Y Pos chromosome, genomic DNAs from female progeny of C57BL/6J Y Pos males contained variable amounts of
### A. Major Y-Chromosome Sequences

<table>
<thead>
<tr>
<th>Name</th>
<th>Source</th>
<th>EcoR I</th>
<th>Hae III, Alu I</th>
<th>Copies / Chrom.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>pBC10-0.6</td>
<td>0.6</td>
<td>~90</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>pBC15-1.1</td>
<td>4.3, 3.7</td>
<td>~60, 30</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>pBA33-1.8</td>
<td>6.7, 1.8</td>
<td>~100 total</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>pBC10-0.17</td>
<td>0.169</td>
<td>~90</td>
<td></td>
</tr>
<tr>
<td><strong>Sxr</strong></td>
<td>[GATA]$_5$</td>
<td>&gt;100, ~23</td>
<td>&gt;1,000 [GATA]$_5$</td>
<td></td>
</tr>
</tbody>
</table>

### B. Y-Specific Fragments

- **Type of Mouse DNA:** Inbred Females (1-10)

- **Y-Chromosome (~40,000 kb):**
  - **Sex Determining Region:** TDF, Sry, Sxr, and Tdy
  - **Centromere**
  - **Pseudoautosomal Region**

- **Y-Specific Fragments:***
  - Lanes: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11
  - Sizes: 2.2, 1.7, 1.4, 1.2, 0.8, 0.6, ***++***
**Fig. 1.**—Locations on Y chromosome and characterization of Y chromosome–linked probes. A, Schematic representation of mouse Y chromosome. The shaded areas, the hatched oval, and the blackened box represent the region of the Y chromosome over which X-Y interchromosomal recombination is suppressed. The superimposed hatched oval, the unblackened box, and the blackened box, represent, respectively, the centromeric, pseudoautosomal, and Sxr regions. Listed above the Y chromosome are the Y-linked probes A–D and (GATA)_3, with their locations indicated by the shading and the blackened box, respectively. Sizes and estimated copy numbers of major Y-specific restriction fragments hybridizing to these probes are also listed above the Y chromosome. Fragments hybridizing to probe C include RFLPs characteristic of Y chromosomes from inbred strains. B and C, EcoRI-digested genomic DNAs (25 µg) from 10 females (lanes 1–10) and 1 male (lane 11) of the inbred strain C57BL/6J, hybridized with replacement synthesized insert of probe A (panel B). EcoRI-digested genomic DNAs (15 µg and 30 µg) from inbred strain females, simultaneously hybridized with nick-translated plasmids containing probe A sequences and the 3.2-kb SalI fragment of α-globin as inserts (panel C, lanes 1 and 2). Autoradiographs were exposed to high-intensity (panels B and C, lanes 1–9) and low-intensity (panel B, lanes 10 and 11; and panel C, lane 10) levels. Arrows identify identical restriction fragments that were electrophoresed on different gels. Major Y-specific EcoRI fragments and their estimated 100 copies are represented by triple-plus signs (+ + +) when present and by a minus sign (−) when absent, below the lanes. D, EcoRI-digested genomic DNAs (25 µg) from 11 females of the inbred strain C57BL/6J (lanes 1–11) and 1 feral male (M; lanes 12 and 13), hybridized with nick-translated insert of probe B. Autoradiographs were exposed to high levels (lanes 1–12). Lane 13 is a low-level exposure of lane 12. Presence and absence of major Y-specific fragments of expected size and copy number are represented as in panels A and B. E, Filters represented in lanes 1–9 of panel D, stripped of probe and autoradiographed to ensure its complete removal. Both DNA on stripped filters and EcoRI-digested feral male (M) DNA on another filter were hybridized with the nick-translated insert of probe C (lanes 1–10). Autoradiographs were exposed to high-intensity (lanes 1–9) and low-intensity (lane 10) levels. The presence and absence of major Y-specific EcoRI fragments of expected 6.6-kb size and copy numbers are represented as in panels A and B.
Y\textsuperscript{pos}-specific sequences. They were detected by hybridization of DNAs with probes A–D and (GATA)_5.

The types of progeny resulting from the background on which the Y\textsuperscript{pos} chromosome was either structurally and functionally unstable (C57BL/6J) or stable (129/Sv) are schematically represented in figure 2. Four types of sibling C57BL/6J Y\textsuperscript{pos} females were classified on the basis of the estimated copy numbers of their Y\textsuperscript{pos} specific sequence content relative to normal males. For simplification of analysis, the major Y\textsuperscript{pos}-specific sequence content of normal C57BL/6J Y\textsuperscript{pos} and 129/Sv Y\textsuperscript{pos} males hybridizing to probes A–D and (GATA)_5 was rounded to 100 copies, or 100%. The four types of C57BL/6J Y\textsuperscript{pos} females identified were High Copy Y, Low Copy Y, Traces Y, and Normal females, which had, respectively, equal amounts, 10%, 1%, and 0% of the Y\textsuperscript{pos}-specific sequence content of males (fig. 2). Some of the phenotypic and karyotypic characteristics of these females are also listed (fig. 2).

Detection of unique-copy levels or traces of Y\textsuperscript{pos}-specific sequences in female DNA with any one probe established Traces Y identity for that female. Regardless of the presence or absence of Y\textsuperscript{pos}-specific sequences hybridizing to the five probes in these analyses, these progeny are likely to be distinct from inbred strains. One example of their difference was the elevated frequency of transmission of a 4.7-kb EcoRI fragment cross-hybridizing with probe B. It was transmitted at six-fold higher frequencies to daughters of C57BL/6J Y\textsuperscript{pos} males than to daughters of inbred strain C57BL/6J males (author's unpublished data). Molecular and genetic analysis of Traces Y and High Copy Y females are presented in the present paper.

Levels of Y\textsuperscript{pos}-specific Sequences Found in Traces Y Females

Transfer of unique-copy levels of Y\textsuperscript{pos}-specific sequences to autosomes or X chromosomes during meiosis is one result predicted by structural instability of the Y\textsuperscript{pos} chromosome. Genomic DNAs from sibling female progeny of C57BL/6J Y\textsuperscript{pos}
males were tested for the presence of $Y^{pos}$-specific sequences by hybridization with probes A–C and (GATA)$_5$. Traces Y females and Normal females that were identified by this analysis are represented by T and F below their respective lanes in figure 3. XX$^{Str}$ and Normal male DNAs are represented by S and M below their respective lanes in figure 3. Estimated levels of $Y^{pos}$-specific restriction fragments identified in each DNA are listed below the respective lanes in figure 3.

$Y^{pos}$-specific EcoRI fragments hybridizing to probe A were of expected size (0.6 kb) and copy number (++++) in male DNAs (fig. 3A, lanes 5 and 6) and absent (−) inNormal female DNAs (fig. 3A, lanes 3 and 4). X-linked internal standard EcoRI fragments were detected in all lanes (fig. 3A, lanes 1–6). The lower levels of the internal standards in male DNAs relative to female DNAs reflected the fivefold lower amounts of male DNAs that were loaded on the gel. Probe A hybridized to ~1.0 copy (+; see arrow in fig. 3A, lane 1) and ~0.1 copies (1/++; see arrow in fig. 3A, lane 2) of the $Y^{pos}$-specific 0.6-kb EcoRI fragment in two female DNAs and therefore identified them as Traces Y DNAs.

$Y^{pos}$-specific EcoRI fragments hybridizing to probe B were of expected size (4.3 kb, 3.7 kb, and 2.2 kb) and copy numbers (++++) in male DNAs (fig. 3B, lanes 5–8) and were absent (−) in Normal female DNAs (fig. 3B, lanes 1, 2, and 4). The internal standard 3.9-kb EcoRI fragment for probe B was detectable in all female DNAs (fig. 3B, lanes 1–4). At the high exposure levels required to detect the 3.9-kb EcoRI fragment, $Y^{pos}$-specific EcoRI fragments in male DNAs hybridized to probe B were obscured. They were clearly visualized at lower exposure levels (fig. 3B, lanes 7 and 8). Probe B hybridized to unique-copy levels of the $Y^{pos}$-specific 4.3-kb EcoRI fragment in one female DNA (+; see arrow in fig. 3B, lane 3) and therefore identified it as Traces Y DNA.

$Y^{pos}$-specific EcoRI fragments hybridizing to probe C were of expected size (6.6 kb) and copy number (+) in mixtures of male/female DNAs (1/250) (fig. 3C, lane 9) and were absent (−) in normal female DNAs (fig. 3C, lanes 7 and 8). The internal standard 3.5-kb EcoRI fragment was detectable in all DNAs (fig. 3C, lanes 1 to 9). Probe C hybridized to unique-copy levels of the $Y^{pos}$-specific 6.6-kb EcoRI fragment in six female DNAs (+; see thick arrow in fig. 3C, lanes 1–6) and therefore identified them as Traces Y DNAs. A minor $Y^{pos}$-specific 9.7-kb EcoRI fragment of probe C was also detected in one Traces Y DNA (fig. 3C, lane 1).

HaeIII-digested DNA from females (fig. 3D, lanes 1–4 and 7–10), XX$^{Str}$ males (fig. 3D, lanes 5 and 11), and normal males (fig. 3D, lanes 6 and 12) were hybridized with end-labeled (GATA)$_5$ oligonucleotide. Autoradiographs of the same filter were exposed to high (fig. 3D, lanes 1–6) and low (fig. 3D, lanes 7–12) intensity levels. HaeIII fragments hybridizing to (GATA)$_5$ and specific for the Sxr region were of expected size (~23 kb and ≥100 kb) and total copy number (>1,000) in XX$^{Str}$ male DNAs (fig. 3D, lanes 5 and 11) and were absent in two Normal female DNAs (fig. 3D, lanes 1, 2, 7, and 8). Significantly lower copy-number levels of the ≥100-kb and ~23-kb HaeIII fragments hybridizing to (GATA)$_5$ were present in Normal male DNA (fig. 3D, lanes 6 and 12). The internal standard HaeIII fragments for (GATA)$_5$, ranging from ~5.0 kb to ~0.5 kb, were detected in all DNAs (fig. 3D, lanes 1–6 and 7–12). Relative to the total levels of ≥100-kb and ~23-kb HaeIII fragments in XX$^{Str}$ male DNA, the (GATA)$_5$ probe hybridized to unique-copy levels of the ~23-kb HaeIII fragment in two female DNAs (see arrow in fig. 3D, lanes 3 and 4) and therefore identified them as Traces Y DNAs. At the high levels of exposure required to detect these ~23-kb HaeIII fragments in Traces Y female DNAs, comigrating fragments in
FIG. 3.—Y\textsuperscript{pos}-specific sequence content of Traces Y female (T) DNAs, shown to be 1% of that of male DNAs. Genomic DNAs were extracted from phenotypically Normal female (F) and male (M) offspring of C57BL/6J × C57BL/6J Y\textsuperscript{pos} × matings. Restriction-endonuclease digests and probes to which digested DNAs were hybridized are specified in panels A–E. Y\textsuperscript{pos}-specific restriction fragments and internal standards for each probe are specified in the text. Thick blackened arrows and unblackened arrows (in conjunction with individually specified lanes) represent, respectively, the presence and absence of major Y\textsuperscript{pos}-specific restriction fragments. The thin blackened arrows are used as in fig. 1. Autoradiographs were densitometrically quantitated and normalized to unique-copy-level internal standards identified in the text. Y\textsuperscript{pos}-specific sequences are represented, by symbols below the lanes, as estimated percentages or copies found in Normal males: +++ = 100% (100 copies); + = 10% (10 copies); +/− = 1% (1 copy); 1/− + = 0.1% (0.1 copy); and − = 0%. Normal and Traces Y females and males identified by this analysis are represented, respectively, as F, T, and M, below the lanes. DNAs in lanes 1, 3, and 4 of panel A are, respectively, from the same females as are the DNAs in lanes 1, 2, and 3 of panel C. DNAs in lanes 1 and 2 of panel E\textsubscript{1} are the same as DNAs in lanes 4 and 5 of panel E\textsubscript{2} and in lanes 7, 8, 10, and 11 of panel E\textsubscript{3}, A, EcoRI-digested genomic DNAs from females (25 µg; lanes 1–4) and males (5 µg; lanes 5 and 6), coelectrophoresed and hybridized with the replacement synthesized insert of probe A. Autoradiographs were exposed to high-intensity levels. B, EcoRI-digested genomic DNAs (25 µg) from females (lanes 1–4) and males (lanes 5–8), coelectrophoresed and hybridized with nick-translated insert of probe B. Lanes 1–6 were exposed to high-intensity levels; lanes 7 and 8 are low-level exposures of lanes 5 and 6. Y\textsuperscript{pos}-specific EcoRI fragments of expected size, as well as copy numbers for probe B, are
Fig. 3 (Continued) bracketed and listed to the right of the figure. The internal standard EcoRI fragment of 3.9 kb for probe B is identified to the left of the figure. C, Female genomic DNAs (25 μg; lanes 1–8) and mixtures of male/female genomic DNAs (0.1 μg/25 μg; lane 9), digested with EcoRI and hybridized with nick-translated insert of probe C. DNAs in lanes 1–9 were coelectrophoresed. D, Female DNAs (15 μg, lanes 1–4 and 7–10), pooled XX<sup>Str</sup> male (S) DNAs (15 μg, lanes 5 and 11), and male DNA (15 μg; lanes 6 and 12), digested with HaeIII, coelectrophoresed, and hybridized with end-labeled (GATA)<sub>3</sub> oligonucleotide. The same filter was autoradiographed to high-intensity (3 d at room temperature; lanes 1–6) and low-intensity (6–8 h at room temperature; lanes 7–12) levels. E, EcoRI-digested genomic DNAs (15 μg) from females (lanes 1 and 2) and a male (lane 3), hybridized with replacement synthesized insert of probe A. Both the filter autoradiographed in lanes 1 and 2 and the filter of male DNA (lane 3) hybridized with probe A were stripped of probe and rehybridized with nick-translated insert of probe C (lanes 4–6). Persistent residual signals, internal standards, and Y-specific EcoRI fragments are identified in the text. Genomic DNAs from the Traces Y female identified above (lanes 1 and 4) was digested with HaeIII. HaeIII-digested genomic DNAs (15 μg) from this Traces Y female (lanes 7 and 10), a Normal female (lanes 8 and 11), and a male (lanes 9 and 12) were coelectrophoresed and hybridized with end-labeled (GATA)<sub>3</sub> oligonucleotide. The same filter was autoradiographed to low-intensity (lanes 7–9) and high-intensity (lanes 10–12) levels of exposure.
XX<sup>Stv</sup> male and normal male (fig. 3D, lanes 5 and 6) DNAs were obscured. Both ≥100-kb and ~23-kb HaeIII fragments were clearly visualized at lower levels of exposure of this autoradiograph (fig. 3D, lanes 11 and 12). Comparison of the internal standard HaeIII fragments of (GATA)<sub>5</sub> in Normal female DNAs (fig. 3D, lanes 1, 2, 7, and 8) and in Traces Y female DNAs (fig. 3D, lanes 3, 4, 9, and 10) showed that all restriction-endonuclease digestions were complete and that more Normal female DNAs than Traces Y female DNAs were loaded on the gel. These results excluded trivial technical explanations such as partial digests and underloading, for the ~23-kb HaeIII fragment of (GATA)<sub>5</sub> being present in Traces Y female but absent in Normal female DNAs.

It was conceivable that, despite precautions, contamination by either plasmid or male genomic DNAs may be trivial technical explanations for the presence of traces of Y-specific sequences in female DNAs. Both of these potential explanations were ruled out by the data presented here. Some, but not all, of the Y-linked probes identified Y<sup>pos</sup>-specific sequences in genomic DNA from the same Traces Y female. Genomic DNAs from the individual Traces Y females represented in lanes 1, 3, and 4 in figure 3A were, respectively, represented in lanes 1, 2, and 3 in figure 3C. Both probe A (fig. 3A) and probe C (fig. 3C) identified Y<sup>pos</sup>-specific sequences in one of these females (fig. 3A, lane 1, and fig. 3C, lane 1). Although the respective internal standards of both probes A and C were clearly visible, only probe C revealed Y<sup>pos</sup>-specific sequences in two of these females (compare fig. 3A, lanes 3 and 4, with fig. 3C, lanes 2 and 3). To further exclude contamination as a potential explanation, Traces Y genomic DNA on the same filter was successively hybridized with probes A and C. Genomic DNA from the same individual Traces Y female hybridized with probes A, C, and (GATA)<sub>5</sub> are represented in lanes 1, 4, 7, and 10 in figure 3E. Y<sup>pos</sup>-specific sequences (see thick blackened arrow in fig. 3E) of probes A (fig. 3E, lanes 1 and 3) and (GATA)<sub>5</sub> (fig. 3E, lanes 7 and 10) of expected size and copy number were detected in this Traces Y female DNA. The filter represented in lanes 1 and 2 of figure 3E was stripped of probe, and both it and a similarly treated filter of male genomic DNA (fig. 3E, lane 3) were rehybridized with probe C (fig. 3E, lanes 4–6). Although the internal standards of probe C (3.5 kb) and probe A (residual signals at 2.2 kb and 1.7 kb) were clearly detected, the expected Y<sup>pos</sup>-specific 6.6-kb EcoRI fragment hybridizing to probe C was absent in this Traces Y female DNA (see unblackened arrow in fig. 3E, lanes 4 and 6). Furthermore, the ~23-kb HaeIII fragment hybridizing to (GATA)<sub>5</sub> was present at unique-copy levels in this Traces Y female DNA (see thick blackened arrow in fig. 3E, lane 10) and was absent in Normal female DNA (fig. 3E, lane 11). The expected copy number of the ~23-kb HaeIII fragment was detected in coelectrophoresed male DNA (fig. 3E, lanes 9 and 12). Contamination by male DNA cannot explain detection of Y<sup>pos</sup>-specific sequences in the same Traces Y DNA by only two of the three Y-linked probes. Since we have no plasmids of ~23-kb size, detection of ~23-kb HaeIII fragments in Y<sup>pos</sup>-specific sequences hybridized with the (GATA)<sub>5</sub> oligonucleotide cannot be attributed to contamination by plasmids. In any case, most plasmid and other DNAs would be digested to small fragments, because of the high frequency of sites for HaeIII. This result also suggested the random representation of loci from the Y<sup>pos</sup> chromosome hybridizing to probes A–C and (GATA)<sub>5</sub> in Y<sup>pos</sup>-specific sequences of Traces Y females. Sequences from the entire length of Y<sup>pos</sup> chromosomes were transmitted to 29%–83% of Traces Y female progeny per generation (table 1; Nallaseth 1987).

Karyotypic analysis of five independent Traces Y females that had 40 chromo-
somes failed to reveal a Y<sup>pos</sup> chromosome, confirming that they were karyotypically XX. The presence of a Y<sup>pos</sup> chromosome was confirmed in two of the male siblings of these females. In an identical mating, Eicher et al. (1982) reported 42 XX females.

Y<sup>pos</sup> chromosomes from High Copy Y Females of N4 and N5 Generations

Sterile sex-reversed, karyotypically XY<sup>pos</sup> (High Copy Y) females represent high proportions of N3 progeny of C57BL/6J Y<sup>pos</sup> males (Eicher et al. 1982; Nallaseth 1987). However, extensive karyotypic analysis failed to reveal gross structural changes of the Y<sup>pos</sup> chromosome from these High Copy Y females (Eicher et al. 1982; author's unpublished data). Mutations due to higher sequence instability in the testis-determining Sxr region, which represents only 0.05% of the Y chromosome, may explain this discrepancy. Although the testis-determining Tdy locus has been convincingly localized to the Sxr region of the murine Y chromosome (Cattanach et al. 1982; McLaren and Monk 1982; Singh and Jones 1982; McLaren et al. 1988; Roberts et al. 1988), fragments hybridizing specifically to the murine Tdy locus have not been cloned (Page et al. 1987; Koopman et al. 1989; Palmer et al. 1989). An alternative approach tested for higher instability of repeated sequences in Sxr regions than in non-Sxr regions of inactive Y<sup>pos</sup> chromosomes. Copy numbers and ratios of restriction fragments from Sxr regions and non-Sxr regions were compared by hybridizing High Copy Y female DNAs with each of probes (GATA)<sub>5</sub> and A–D.

EcoRI-digested genomic DNAs from High Copy Y (fig. 4A, lanes 1 and 2) and Normal females (fig. 4A, lane 3) and from inbred strain (fig. 4A, lane 4), 129/Sv Y<sup>pos</sup> (fig. 4A, lane 5), and C57BL/6J Y<sup>pos</sup> (fig. 4A, lane 6) males were hybridized with probes A–C and rDNA. As expected, the rDNA-specific and Y<sup>pos</sup>-specific EcoRI fragments were, respectively, present and absent in Normal female DNA. The multicyclop 1.8-kb EcoRI fragment hybridizing to probe C was an RFLP that was characteristic of inbred strain (C57BL/6J) Y chromosomes. It was present at unique-copy levels in DNA from males with the Y<sup>pos</sup> chromosome and in DNA from High Copy Y females. This confirmed the identity of the Y<sup>pos</sup> chromosome in these females. In these High Copy Y female and male genomic DNAs, there were no differences in the mobilities of major EcoRI fragments hybridizing to probes A–C and rDNA (fig. 4A). Y<sup>pos</sup>-specific EcoRI fragments from 31 High Copy Y female DNAs and from one of each of the three types of listed male DNAs were densitometrically quantitated, and their ratios were plotted as per the legend to figure 4C. Values of hybridization ratios of seven pairs of Y<sup>pos</sup>-specific EcoRI fragments hybridizing to probes A–C (fig. 4A and fig. 4C, lower panel) did not differ significantly (>2.5 X) between either male or inbred-strain-male DNAs and the mean for High Copy Y female DNAs (fig. 4C, upper panel). Furthermore, in High Copy Y female DNAs, the range and the mean of the hybridization ratios of each of these seven pairs of Y<sup>pos</sup>-specific EcoRI fragments also did not differ significantly (>2.5 X) (fig. 4C, lower panel).

Probe D, however, which also hybridized to sequences from the non-Sxr region, identified maximal sequence instability of Y<sup>pos</sup> chromosomes. All ~100 copies of Y<sup>pos</sup>-specific EcoRI fragments (fig. 5B, lanes 1 and 5) hybridizing to probe D shifted from 169 bp (see unblackened arrow in fig. 5B, lane 2) to 225 bp (see blackened arrow in fig. 5B, lane 2) in one High Copy Y female DNA. Only partial shifts were detected in High Copy Y female sibling DNAs (fig. 5B, lanes 2–4), in male DNAs (fig. 5B, lane 5), and in 27 other High Copy Y female DNAs, also from N4 and N5 generations.
(data not shown). To accentuate instability of sequences hybridizing to probe D, only genomic DNAs displaying maximal 225-bp/169-bp hybridization ratios were densitometrically quantitated and plotted as per the legend to figure 4C. The mean of the hybridization ratios for probe D fragments in High Copy Y female DNAs (≥8.9) was 11–22-fold higher than their hybridization ratios in male DNAs (0.4 and 0.8) (fig. 4C, upper panel). In High Copy Y female DNAs the hybridization ratios for probe D fragments was 0.1–26, or 260-fold, which was 29-fold higher than their mean hybridization ratios (≈8.9) (fig. 4C, lower panel).

However, only one of eight pairs of Y^Pos-specific EcoRI fragments from the non-Sxr region displayed this high degree of instability. The remaining seven pairs of Y^Pos-specific EcoRI fragments were stable in 31 High Copy Y female DNAs from three types of tissue (kidney, liver, and spleen) and from two backcross generations. Thus, non-Sxr regions of inactive Y^Pos chromosomes from High Copy Y females of N4 and N5 generations displayed great stability during both meiotic and somatic cell division.

In contrast, the Sxr regions of inactive Y^Pos chromosomes from the aforementioned High Copy Y females were significantly unstable. Genomic DNAs from these High Copy Y females (fig. 4B, lanes 1, 2, 4, and 5), a Normal female (fig. 4B, lane 3), a Normal N3 male (fig. 4B, lane 7), and XX^Sxr males (fig. 4B, lane 6) were digested with either HaeIII (fig. 4B) or AluI (data not shown) and hybridized with the end-labeled eicosanucleotide (GATA)^5. Internal standards showed that digests were complete, and equal amounts of DNA were loaded. Significant variations in copy numbers and hybridization ratios of the ≥100-kb/≈23-kb HaeIII (fig. 4B) and AluI (data not shown) fragments hybridizing to (GATA)^5 were detected within High Copy Y female DNAs and between High Copy Y female, male, and pooled XX^Sxr male DNAs (see arrows in fig. 4B, lanes 1–7). The ≥100-kb and ≈23-kb HaeIII and AluI fragments hybridizing to (GATA)^5 were densitometrically quantitated and plotted as per the legend to figure 4C. XX^Sxr males have only ≈0.05% of the Y chromosome including the Tdy locus (Cattanach et al. 1982; Evans and Burtneshaw 1982; Singh and Jones 1982). However, in pooled XX^Sxr male genomic DNAs, the copy number of the ≥100-kb HaeIII fragments was at least 11-fold higher than that in genomic DNAs from High Copy Y females (fig. 4B, lanes 1, 2, 4, and 5) and males (fig. 4B, lane 7). The mean of the hybridization ratios of ≥100-kb/≈23-kb HaeIII and AluI fragments hybridizing to (GATA)^5 in High Copy Y female genomic DNAs was ≈3.7 (fig. 4C, upper panel). It did not differ significantly from the hybridization ratios of the ≥100-kb/≈23-kb HaeIII fragments of (GATA)^5, either in pooled XX^Sxr male genomic DNAs (≈3.8) or in genomic DNAs from an N3 male (≈1.97) (fig. 4C, upper panel). In contrast, in High Copy Y female genomic DNAs, the hybridization ratios of ≥100-kb/≈23-kb HaeIII and AluI fragments of (GATA)^5 were 166.5 and 100, respectively (fig. 4C, lower panel). These ranges were, respectively, 45-fold and 51-fold higher than the mean hybridization ratios of these HaeIII and AluI fragments in High Copy Y female genomic DNAs. Phenotypic inactivity of Sxr regions of Y^Pos chromosomes in High Copy Y females correlates well with the molecular instability of their repeated sequence content. It may be that the high variability and the copy-number reductions of sequences in the Sxr regions relative to sequences in the non-Sxr regions reflect the indirect selection for analysis of inactive Y^Pos chromosomes, i.e., High Copy Y females.
Instability and Inactivation of Murine Y Chromosomes

Hybridization Ratio (range)

Hybridization Ratio (Mean)

FIG. 4.—High Copy Y females (HC) of N4 and N5 backcross generations whose Y<sup>Poa</sup> chromosomes were more extensively altered in Sxr regions than in non-Sxr regions. Types of matings and identities of symbols or abbreviations are as in the legend to fig. 3. A, EcoRI-digested genomic DNAs (20 μg) from High Copy Y females (lanes 1 and 2), a Normal female (lane 3), and males (lanes 4–6), coelectrophoresed and hybridized with probes A–C and with ribosomal DNA (rDNA) as internal standard. High Copy Y females were from the N4 and N5 backcross generations. Males were of the inbred strain C57BL/6J (lane 4), 129/Sv Y<sup>Poa</sup> (lane 5), and C57BL/6J Y<sup>Poa</sup> (lane 6). Major Y-specific EcoRI fragment sizes—and probes to which they hybridized—are listed to the right. Autoradiographs were exposed to intermediate-intensity levels. Lower exposures of autoradiographs with a linear range of signals were densitometrically quantitated to provide the data plotted in panel C. B, HaeIII-digested genomic DNAs (10 μg) from some of High Copy Y females (lanes 1, 2, 4, and 5), Normal females (lane 3), XX<sup>Δ+</sup> males (pooled DNAs; lane 6), and a Normal N3 male (lane 7), hybridized with end-labeled (GATA)<sub>5</sub> oligonucleotide. DNAs in lanes 1 and 2, lanes 3 and 5, and lanes 6 and 7 were coelectrophoresed. Autoradiographs were exposed to low-intensity levels (6–8 h at room temperature). These and other autoradiographs were densitometrically quantitated for the data plotted in panel C. C, rDNA-specific and Y<sup>Poa</sup>-specific EcoRI fragments of rDNA, probes A–C (panel A and fig. 5A) and probe D (fig. 5B), densitometrically scanned. Areas below peaks were integrated, and the area for each Y<sup>Poa</sup>-specific EcoRI fragment was normalized to the area for the 12-kb EcoRI fragment of rDNA (internal standard). Sxr region specific ≥100 kb and ~23 kb HaeIII (panel B) and AluI (data not shown) fragments hybridizing to (GATA)<sub>5</sub> were also densitometrically scanned. Areas below these two peaks were integrated and normalized to integrated areas below the fragments from ~5.0 kb to ~0.5 kb, which served as internal standards for the (GATA)<sub>5</sub> probe. Hybridization ratios of pairs of restriction fragments identified on the X-axis of the lower graph of panel C were calculated for each of the genomic DNAs identified in the upper graph of panel C. The mean of hybridization ratios of pairs of fragments was determined for High Copy Y female DNAs and for all male DNAs. The range from minimal to maximal hybridization ratios of pairs of fragments was determined for High Copy Y female DNAs. The polymorphic 1.8-kb EcoRI fragment hybridizing to probe C was only quantitated in DNA from inbred-strain males. Probes A–D were absent in XX<sup>Δ+</sup> male DNA. All DNAs, unless stated otherwise, were extracted from livers. Probes A–D were hybridized to 31 DNAs from N4 and N5 HC-Y female siblings and to 1 DNA from an N3 male parent; three types of tissues (kidney, liver, and spleen) from each of eight High Copy Y females plus one type of tissue from seven High Copy Y females totaled 31 DNAs. (GATA)<sub>5</sub> was hybridized to 15 DNAs from N4 and N5 High Copy Y female siblings and from 1 DNA from an N3 male parent; two types of tissue (kidney and liver) from each of seven High Copy Y females plus one type of tissue from 1 High Copy Y female totaled 15 High Copy Y DNAs. <sup>[6]</sup>C57BL/6J (XY<sup>B6</sup>) <sup>[8]</sup>; <sup>[6]</sup>C57BL6J Y<sup>Poa</sup> (XY<sup>Poa</sup>) <sup>[8]</sup>; <sup>[6]</sup>129/Sv Y<sup>Poa</sup> (XY<sup>Poa</sup>) <sup>[8]</sup>; <sup>[6]</sup>Sxr (XX<sup>Δ+</sup>) <sup>[8]</sup>; and <sup>[6]</sup>= High Copy Y (XY<sup>Poa</sup>) <sup>5</sup>.

Restriction Analysis: Exclusion of Mechanisms of Extrachromosomal Transmission of Y<sup>Poa</sup>-specific Sequences

One explanation for the origin of Y<sup>Poa</sup>-specific sequences in Low Copy Y and Traces Y females is extrachromosomal transmission via circular elements, retrotransposons, or transposons. One common feature to all three of these mechanisms of
extrachromosomal transmission is the random loss or gain of restriction fragments at the sites of excision and/or reintegration, at high frequencies.

More than 90% of High Copy Y, Low Copy Y, and Traces Y females failed to reveal Y Pos-specific EcoRI fragments of unexpected sizes when hybridized with probes A–C. Those High Copy Y and Traces Y female DNAs that displayed EcoRI fragments of expected size were further analyzed with different enzymes or their various combinations, so that there were a total of 17 distinct types of digests. Digests were designed to either release or retain genomic sequences flanking Y Pos-specific sequences hybrid-
izing to each probe in High Copy Y, Low Copy Y, and Traces Y females. Enzymes with four-base and six-base recognition sites were included (Nallaseth 1987). Randomization of restriction fragments characteristic of induction of retroviruses (Jenkins and Copeland 1985) or transposons (Bingham et al. 1982) and excision into extrachromosomal sequences (Smith and Vinograd 1972; Stanfield and Lengyel 1979; Calabreta et al. 1982; Krolewski et al. 1982; Potter 1984; Walbot and Cullis 1985) were not detected. Thus, it was concluded that these mechanisms do not account for generation of Y pos-specific sequences in these three types of females. Some of the remaining ~10% of these female DNAs that displayed Y pos-specific EcoRI fragments of unexpected sizes that hybridized to probes A-C were analyzed below.

Two distinct mechanisms for losses of Y pos-specific EcoRI fragments were defined by Traces Y and High Copy Y female DNAs. One of the effects on High Copy Y DNA further excluded extrachromosomal transmission of Y pos-specific sequences as a possible mechanism.

Major Y pos-specific EcoRI fragments of expected size (0.6 kb) and copy number (~100) were present (++++) in M. poschiavinus (Valende) male genomic DNA hybridized with probe A (fig. 6, lane 7). Characteristic minor Y pos-specific EcoRI fragments (4.2 kb and 3.8 kb) were also present at ~10-fold lower copy number in M. poschiavinus (Valende) male genomic DNA hybridized with probe A. Both major and minor Y pos-specific EcoRI fragments hybridizing to probe A remained absent (−) from Normal female genomic DNAs, even at high exposures (fig. 6, lanes 1 and 5). Internal standard EcoRI fragments (2.2 kb, 1.7 kb, and 0.8 kb) hybridizing to probe A were present in all DNAs and showed that restriction-endonuclease digestions of genomic DNAs were complete. Y pos-specific EcoRI fragments of the expected size of 0.6 kb were detected in two Traces Y female genomic DNAs hybridized with probe A (see thick blackened arrows in fig. 6, lanes 4 and 6). However, in genomic DNAs from two Traces Y female siblings hybridized with probe A, an EcoRI fragment of the unexpected size of 5.4 kb was present at unique-copy levels (see blackened arrows in fig. 6, lanes 2 and 3). Both 0.6-kb and 5.4-kb EcoRI fragments were present in genomic DNA from a third Traces Y female hybridized with probe A (see blackened arrows in fig. 6, lane 4). Overexposure of autoradiographs of M. poschiavinus (Valende) male genomic DNA hybridized with probe A failed to reveal a similar sized (5.4-kb) EcoRI fragment (data not shown). Therefore, the 5.4-kb EcoRI fragments were reproducibly generated during transfer of Y pos-specific sequences to autosomes/X chromosomes transmitted to two Traces Y female siblings (fig. 6, lanes 2 and 3). Because double crossovers are rare events, at least one of the restriction sites flanking the 0.6-kb EcoRI fragment hybridizing to probe A must have been lost during transfer of Y pos-specific sequences.

In contrast to the above loss of restriction sites of a single copy fragment, precise mobility shifts of all ~100 copies of Y pos-specific restriction fragments were detected in genomic DNAs from some High Copy Y females. These High Copy Y females were also from the N5 backcross generation. The multicopy Y pos-specific sequences hybridizing to probes A and D were well defined by restriction endonuclease mapping and sequencing. The Y pos-specific 600-bp EcoRI fragment of probe A and the Y pos-specific 169-bp EcoRI fragment of probe D were contiguous and present at the frequency of once per ~14-kb HpaI repeat unit. Thus, these two probes hybridized to sequences spanning (~14 kb × ~100 copies ≈) 1,400 kb of the Y pos chromosome. Approximately 50% of the sequences within the 169-bp EcoRI fragment of probe D
Fig. 6.—Reproducible losses of EcoRI restriction sites of Y^{pos}-specific sequences in Traces Y female DNAs. Types of matings and identities of symbols and abbreviations are as in fig. 3. EcoRI-digested genomic DNAs (25 μg) from Normal females (lanes 1 and 5), Traces Y females (lanes 2–4 and 6), and a Mus poschiavinus (Valende) male (lane 7) were hybridized with the replacement synthesized insert of probe A. The DNAs in lanes 1–3 and in lanes 4–6 were extracted from siblings and were coelectrophoresed. Autoradiographs were exposed to high-intensity (lanes 1–6) and low-intensity (lane 7) levels. Sizes of EcoRI fragments hybridizing to probe A are listed to the right.

consists of the 3' terminus of rodent Alu (B1) sequences and their oligo A tail (Nallaseth 1987).

Expected sizes and copy numbers of EcoRI fragments hybridizing to probes A–C and rDNA were present in genomic DNAs from three High Copy Y female siblings (fig. 5A, lanes 2–4) and from males (fig. 5A, lane 5). Y^{pos}-specific EcoRI fragments were absent from Normal female DNAs (fig. 5A, lane 1). The filter was stripped of probes and rehybridized with probe D (fig. 5B). Probe D hybridized to \( \sim 100 \) copies of a 169-bp EcoRI fragment that was present in male DNA (fig. 5, lane 5) but absent in Normal female DNA (fig. 5B, lane 1). In genomic DNA from one High Copy Y female sibling, all \( \sim 100 \) copies of the Y^{pos}-specific EcoRI fragments hybridizing to probe D had precisely shifted from 169 bp (see unblackened arrow in fig. 5B, lane 2) to 225 bp (see blackened arrow in fig. 5B, lane 2). Partial shifts were detected in genomic DNAs of sibling High Copy Y females (fig. 5B, lanes 3 and 4) and in males (fig. 5B, lane 5). Thus, these mobility shifts were due to single-generation events, either during development in the (N4) paternal germ line or in the (N5) somatic cells.
Instability and Inactivation of Murine Y Chromosomes

in High Copy Y females. Precise mobility shifts of the Y\textsuperscript{pos}-specific 169-bp EcoRI fragment occurred at the high frequency of 8 of 48 XY\textsuperscript{pos} offspring from the backcross generations N3–N7. Restriction analysis showed that these mobility shifts were limited to the 169-bp EcoRI fragment within the ~14-kb HpaI repeat unit (author's unpublished data).

Trivial technical explanations such as partial digests or aberrant electrophoresis cannot explain either the reproducibility and precision or the limitation of mobility shifts to one of five multicopy EcoRI fragments within a single lane. The molecular mechanisms causing mobility shifts must operate coordinately, precisely, and selectively on each of ~100 copies of the 169-bp EcoRI fragment interspersed over ≥1,400 kb of the Y\textsuperscript{pos} chromosome. Because all other Y\textsuperscript{pos}-specific repeats were intact and the mobility shifts were limited to 169 bp within the ~14-kb HpaI repeat unit, intrachromosomal events rather than extrachromosomal transmission are the likely mechanisms. These two results showed that deletions and/or rearrangements of Y\textsuperscript{pos} chromosomes were likely to be associated with the transfer of Y\textsuperscript{pos}-specific sequences to the autosomes or X chromosomes that were transmitted to Traces Y females.

Sequences First Deleted and Transmitted from Y\textsuperscript{pos} Chromosomes during Meiosis in F1 Males

The C57BL/6J and 129/Sv genetic backgrounds are, respectively, inductive and noninductive for dysfunction of Y\textsuperscript{pos}-encoded testis determination. A correlation has been detected among the presence in sibling females of traces of Y\textsuperscript{pos}-specific sequences and intact inactive Y\textsuperscript{pos} chromosomes, the selective instability of repeated sequences in the Sxr region, and nonfunctional testis determination. To test whether the aberrant recombinational mechanisms that randomly delete Y\textsuperscript{pos}-specific sequences from Y\textsuperscript{pos} chromosomes actually correlate with the disruption of testis determination, a genetic approach was taken. The Y\textsuperscript{pos} chromosome was backcrossed from the 129/Sv to the C57BL/6J genetic backgrounds, where it was structurally and functionally destabilized at high frequencies. If the transmission of Y\textsuperscript{pos}-specific sequences to XX females (i.e., as deletion products) and the functional inactivity of Y\textsuperscript{pos} chromosomes in their XY\textsuperscript{pos} female siblings result from related mechanisms, then four predictions can be made for successive backcross generations (N). First, traces of Y\textsuperscript{pos}-specific sequences should not be detected on the 129/Sv genetic background on which the testis-determining Sxr region is fully functional. Second, traces of Y\textsuperscript{pos}-specific sequences should not be detected in somatic tissues of F1 hybrids of C57BL/6J x 129/Sv Y\textsuperscript{pos} females. Third, traces of Y\textsuperscript{pos}-specific sequences should first be detected in backcross generations preceding the backcross generation in which inactive Y\textsuperscript{pos} chromosomes of High Copy Y females are detected. Fourth, generating of traces (i.e., deletions) should have a cumulative effect on the functional inactivity of a population of Y\textsuperscript{pos} chromosomes in successive backcross generations. These four predictions of a relationship between the generation of Y\textsuperscript{pos}-specific trace sequences (i.e., deletion products) and the functional inactivation of the Sxr region on the Y\textsuperscript{pos} chromosome were examined and found to be correct. They are addressed below.

To minimize the presence of non-129/Sv sequence content, a male of the 129/Sv Y\textsuperscript{pos} (N18) strain was used to produce (C57BL/6J x 129/Sv Y\textsuperscript{pos}) F1 hybrid offspring. Genomic DNAs from F1 hybrid females were analyzed for their Y\textsuperscript{pos}-specific sequence content, and their sibling F1 hybrid males were mated with pure-strain C57BL/6J females. Genomic DNAs of [C57BL/6J x F1 (C57BL/6J x 129/Sv Y\textsuperscript{pos})] N2 female offspring were analyzed for their Y\textsuperscript{pos}-specific sequence content.
Similarly, a backcross series was independently initiated with a mating of C57BL/6J × C57BL/6J YPos (N3), and genomic DNAs of N4–N7 female offspring were analyzed for their YPos-specific sequence content. The results of these matings are shown in figures 3–8, and they are summarized in table 1.

Genomic DNAs from an inbred strain male (fig. 7A, lane 12, fig. 7B, lane 10) and female (fig. 7A, lane 11, and fig. 7B, lane 9), hybridized with probe A, displayed Y-specific and cross-hybridizing EcoRI fragments of expected sizes and copy numbers. The cross-hybridizing EcoRI fragments of probe A were clearly detected in genomic DNAs from F1 females (fig. 7A, lanes 1–10). However, the YPos-specific 0.6-kb EcoRI fragment hybridizing to probe A was absent from these F1 female DNAs (fig. 7A, lanes 1–10). Similarly, the respective presence and absence of cross-hybridizing and Y-specific EcoRI fragments of probe C in F1 female genomic DNAs was established (data not shown). Therefore, sequences from YPos chromosomes were not transferred to either autosomes or X chromosomes in germ lines of 129/Sv YPos males. Furthermore, cross-hybridizing sequences of probes A and C were not amplified and rearranged so as to acquire electrophoretic mobilities of YPos-specific sequences during the somatic development of F1 progeny.

High frequencies (75%; table 1) of genomic DNAs from N2 females hybridized with probe A (see thick arrow in fig. 7B, lanes 1–8) revealed unique-copy levels of the YPos-specific 0.6-kb EcoRI fragment. Genomic DNAs of these N2 females hybridized with probe C and (GATA)3 also revealed unique-copy levels of YPos-specific fragments (data not shown). Therefore, YPos-specific sequences were first transferred to autosomes or X chromosomes in germ lines of F1 males.

Genetic heterozygosity was sufficient—and predominant proportions of C57BL/6J alleles were therefore not required—for generating traces of YPos-specific sequences (i.e., deletion products). Genetic heterozygosity of C57BL/6J genetic backgrounds was not sufficient for generating High Copy Y females which were first detected in N3 generations. Thus, the second and third predictions of a causal relationship between the generating of YPos-specific deletion products and functional inactivity of YPos chromosomes were observed to be true.

The frequencies of Traces Y females containing YPos-specific sequences in the backcross generations N2, N4, N5, and N6 were, respectively, 75%, 52%, 29%, and 39% (table 1). A χ² analysis (χ² = 50.05) with 1 degree of freedom showed a significant probability (P < 0.001) that these traces of Y-specific sequences would be transmitted to XX females at frequencies compatible with normal pseudoautosomal exchange. In contrast, the frequencies of High Copy Y females in the backcross generations N2, N4, N5, and N6 respectively, were 0%, 14%, 49%, and 31% (table 1). Deleting YPos-specific sequences from a population of YPos chromosomes apparently had a cumulative effect on the inactivation of their testis-determining (Sxr region) functions in subsequent backcross generations. Thus, the fourth prediction of a causal relationship also held true.

The absolute requirement for the germ-line presence of a functional YPos chromosome for the transmission of traces of YPos-specific sequences to female offspring was also established. Female progeny of (C57BL/6J YPos Normal females (N4) × C57BL/6J pure-strain males) matings lacked any YPos-specific sequences (table 1; Nallaseth 1987). The absence of YPos-specific sequences in these Normal female (N5) progeny further excluded the possibility that amplifications and rearrangements of the cross-hybridizing sequences of Y-specific probes would allow them to acquire YPos-specific mobilities.
Fig. 7.—Trace levels of Y-pos-specific sequences first generated during meiotic division in F1 males. Identities of symbols and abbreviations are as in figs. 3 and 4. EcoRI-digested genomic DNAs (25 μg) from (C57BL/6J × 129/Sv Y pos) F1 females (panel A, lanes 1–10), (C57BL/6J × F1) (C57BL/6J × 129/Sv Y pos) N2 females (panel B, lanes 1–8), C57BL/6J pure-strain females and males (panel A, lanes 11 and 12, and panel B, lanes 9 and 10) were hybridized with the replacement synthesized insert of probe A. Genomic DNAs in panel A, lanes 1–10 in panel A, lanes 11 and 12, in panel B, lanes 1–8, and in panel B, lanes 9 and 10, were coelectrophoresed. Sibling males of F1 females were mated with inbred-strain C57BL/6J females to obtain N2 female offspring.
Fig. 8.—Precise structural alterations of stable multicopy Y$^{Pos}$-specific EcoRI fragments in DNAs of High Copy Y females from terminal backcross generations (N6 and N7). Symbols and abbreviations are as in figs. 3 and 4. K, L, and S denote lanes showing DNAs from kidney, liver, and spleen, respectively, of individual High Copy Y females. Sibling males of High Copy Y females of the N5 generation (analyzed in fig. 4A–C) sired High Copy Y and Normal female siblings of the N6 generation (analyzed in this fig.). Sibling males of High Copy Y females (analyzed in this fig.) sired High Copy Y female siblings of the N7 generation (analyzed panel C). All XY$^{Pos}$ offspring in N7 litters were sterile. A and B, EcoRI-digested genomic DNAs (20 μg) from a Normal female (panels A and B, lanes 1) and High Copy Y female siblings from the N6 generation (panels A and B, lanes 2–7) were coelectrophoresed and hybridized either with probe C (panel B) or with rDNA and probes A–C (panel A). Autoradiographs were exposed to low-intensity (panel A) and high-intensity (panel B) levels. Identities and sizes of Y$^{Pos}$-specific and rDNA-specific EcoRI fragments are listed to the left and right, respectively, of panel A and to the right of panel B. C, Genomic DNAs extracted from livers of High Copy Y female siblings of N7 generation. EcoRI-digested genomic DNAs (20 μg) were coelectrophoresed and hybridized with probes A–C. Autoradiographs were exposed to low-intensity levels.

Stable Repeated Sequences on Y$^{Pos}$ Chromosomes Showing Precise Structural Changes in Terminal Backcross Generations (N6 and N7)

The stability of a single population of Y$^{Pos}$ chromosomes was followed over the four backcross generations N4–N7, by hybridization of High Copy Y female genomic DNAs with probes A–D (figs. 4, 6, and 8 and table 1). All XY$^{Pos}$ progeny in the N7 generation were sterile, thus ending this mouse line. Densitometric analysis of genomic DNAs from High Copy Y females of N4 and N5 generations showed that Y$^{Pos}$-specific EcoRI fragments hybridizing to probe D and to probes A–C were, respectively, unstable and stable during germ-line and/or somatic development (figs. 4A and 4C and 6). The mean hybridization ratios of two major Y$^{Pos}$-specific EcoRI fragments (4.3 kb and 3.7 kb) hybridizing to probe B was 2.0 (fig. 4C). This hybridization ratio was
### Table 1
Types of C57BL/6J $Y^\text{pos}$ Females and Their Frequencies/Generation

<table>
<thead>
<tr>
<th>PARENTAL MATING</th>
<th>GENERATION OF OFFSPRING</th>
<th>% OF FEMALE PROGENY</th>
<th>NO. OF FEMALES ASSAYED</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(C57BL/6J $Y^\text{pos}$)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Normal*</td>
<td>Traces</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(F)</td>
<td>Y</td>
</tr>
<tr>
<td>C57BL/6J (P) $\times$ 129/Sv $Y^\text{pos}$ (d)</td>
<td>F1</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>C57BL/6J (P) $\times$ FL $Y^\text{pos}$ (d)</td>
<td>N2</td>
<td>25</td>
<td>75*</td>
</tr>
<tr>
<td>C57BL/6J (P) $\times$ N3 $Y^\text{pos}$ (d)</td>
<td>N4</td>
<td>14*</td>
<td>52</td>
</tr>
<tr>
<td>C57BL/6J (P) $\times$ N4 $Y^\text{pos}$ (d)</td>
<td>N5</td>
<td>17*</td>
<td>29</td>
</tr>
<tr>
<td>C57BL/6J (P) $\times$ N5 $Y^\text{pos}$ (d)</td>
<td>N6</td>
<td>0*</td>
<td>39</td>
</tr>
<tr>
<td>C57BL/6J $Y^\text{pos}$ (F) (N4) (9)</td>
<td>N5</td>
<td>100</td>
<td>0</td>
</tr>
</tbody>
</table>

**NOTE.**—All female offspring of C57BL/6J (Q) $\times$ C57BL/6J $Y^\text{pos}$ (d) matings are classified as C57BL/6J $Y^\text{pos}$ females.

* Progeny in which $Y^\text{pos}$-specific sequences were not detected with the three moderately repeated Y-specific probes denoted as F.

* Not counted in the number of females assayed.

* Three or four probes present at trace levels; fourth probe present at low copy levels.

* Data from Normal (or Traces Y) females in which the presence or absence of $Y^\text{pos}$-specific sequences was not conclusively established are omitted.

* $Y^\text{pos}$ = Y chromosome of Mus musculus on different genetic backgrounds.

---

Invariant (<0.5 ×) in 31 genomic DNAs extracted from three types of tissue (kidney, liver, and spleen) from these High Copy Y females and from males with either Y $^\text{pos}$ chromosomes or inbred strain Y chromosomes.

However, EcoRI-digested genomic DNAs from High Copy Y females of the N6 and N7 generation hybridized with probe B showed that the ratio of the 4.3-kb and 3.7-kb fragments was altered >30-fold. All copies of the major 3.7-kb and minor 2.2-kb Y $^\text{pos}$-specific EcoRI fragments hybridizing to probe B were clearly absent in spleen DNA—but not in kidney or liver DNAs—from one High Copy Y female sibling of the N6 generation (see unblackened arrows in fig. 8A, lane 7). All other major Y $^\text{pos}$-specific EcoRI fragments hybridizing to probes A–C were identical in kidney, liver, and spleen DNAs from both N6 siblings (fig. 8A, lanes 2–6). Losses of these two fragments occurred in one of four and two of three High Copy Y females of the N6 and N7 generations, respectively. Although genomic DNA in lane 2 of figure 8A was smeared, major Y $^\text{pos}$-specific EcoRI fragments hybridizing to probes A–C were visible. In addition, ratios of the rDNA-specific (12-kb) EcoRI fragment to each of the major Y $^\text{pos}$-specific EcoRI fragments of probes A–C differed significantly (i.e., more than fivefold) between genomic DNAs from High Copy Y females of the N6 generation (fig. 8A, lanes 2–6) and N4 and N5 generations (fig. 4A).

The filter that was autoradiographed for figure 8A was previously probed with probe C and exposed to high levels (fig. 8B). Minor Y $^\text{pos}$-specific 180-bp EcoRI fragments hybridizing to probe C were present in liver DNAs (see blackened arrow in fig. 8B, lanes 3 and 6) but were absent in kidney and spleen DNAs (fig. 8B, lanes 2, 4, 5, and 7) from both of these High Copy Y female siblings of the N6 generation.

The filter autoradiographed for figure 8A was stripped of probe and reannealed with probe D. All copies of the Y $^\text{pos}$-specific 169-bp EcoRI fragment hybridizing to probe D displayed mobility shifts from 169 bp to 110 bp in liver DNAs—but not in
kidney and spleen DNAs—from both of these High Copy Y female siblings of the N6 generation (data not shown).

In EcoRI-digested genomic DNAs from livers of two High Copy Y female siblings of the N7 generation, the major 3.7-kb and the minor 2.2-kb fragments hybridizing to probe B were absent in one (see unblackened arrows in fig. 8C, lane 2) but were present in the other (fig. 8C, lane 1). All other major Y Pos-specific EcoRI fragments hybridizing to probes A–C were indistinguishable in DNAs from these two N7 siblings (fig. 8B, lanes 1 and 2).

Therefore, multicopy Y Pos-specific EcoRI fragments hybridizing to each of the three probes B–D underwent distinct types of precise structural changes in different tissues. The Y Pos chromosome was structurally destabilized during the somatic development of High Copy Y females. Trivial technical explanations for these mobility shifts included either partial and star activity of restriction endonucleases or aberrant electrophoretic conditions. Star activity was tested for and was empirically excluded; furthermore, these explanations were incompatible both with the precision and reproducibility of mobility shifts between different genomic DNAs and with their limitation to one or two of five multicopy Y Pos-specific EcoRI fragments of DNA within a single lane.

Therefore, eventually, even stable repeated sequences on Y Pos chromosomes showed precise structural alterations at high frequencies. These Y Pos chromosomes were structurally destabilized in terminal backcross generations, suggesting that their maintenance on C57BL/6J backgrounds had some cumulative effect.

Discussion

Structural Destabilization: Functional Inactivation of Feral Y Pos Chromosomes on C57BL/6J Inbred Backgrounds

The data presented in the present paper demonstrated that the Y Pos chromosome from the feral species Mus poschiavinus was structurally destabilized at high frequencies when it was backcrossed onto the C57BL/6J inbred genetic background. However, the Y Pos chromosome was stable on the 129/Sv inbred genetic background. Molecular and genetic analysis showed (a) that there was a strong correlation between the structural instability of the Y Pos chromosome on C57BL/6J backgrounds and eventual functional inactivation of its Tdy locus and (b) that the Y Pos chromosome functioned normally on 129/Sv genetic backgrounds (fig. 2).

Two results predicted by the structural destabilization of Y Pos chromosomes were confirmed, by using Y-linked multicopy sequences, by probing genomic DNAs from daughters of C57BL/6J Y Pos males. In genomic DNAs from XX females, unique-copy levels or traces of Y Pos-specific sequences were present at high frequencies (29%–75%/Y-chromosomal repeat sequence/generation). These were Traces Y females. In genomic DNAs from some of their XY Pos female siblings, multiple copies of Y Pos-specific restriction fragments either were absent or underwent precise coordinated mobility shifts. These were High Copy Y females with structurally destabilized Y Pos chromosomes. Trivial technical explanations for trace sequences in Traces Y females and for destabilized Y Pos chromosomes in High Copy Y females were excluded. Data for Low Copy Y females are not presented.

In 80 N4 offspring from matings that were identical to those in the present study, Eicher et al. (1982) identified 42 XX females, 16 XY Pos females, and 22 XY Pos with ovaries and/or ovotestis. Five Traces Y and 2 High Copy Y females were karyotypically XX and XY Pos, respectively (author's unpublished data). Given that 11 of 21 N4
females analyzed in the present work were Traces Y females, at least some of the XX females in the study by Eicher et al. (1982) were also Traces Y females. Inability to detect either presence of Ypos chromosomal fragments or gross structural changes of Ypos chromosomes in karyotypes of XX and XYpos females could be explained by two results from this work: (1) in High Copy Y females of the N4 or N5 generation, repeated sequence copy numbers and stabilities were more extensively reduced in $Sx$ regions containing the $Tdy$ locus and representing <0.05% of Ypos chromosomes than in non-$Sx$ regions; and (2) since fragments <200 kb cannot be visualized in optical microscopes, both Traces Y and Low Copy Y females would escape detection.

Although it would be ideal to show that deletions/rearrangements of the $Tdy$ locus inactivate testis-determining functions of Ypos chromosomes, current understanding of this complex developmental pathway (Singh et al. 1984; Chandra 1985; Page et al. 1987; Koopman et al. 1989; Palmer et al. 1989; Schneider-Gädicke et al. 1989) precludes such an unambiguous correlation. Instead, when a genetic approach was used, increased structural instability indirectly correlated with the functional inactivity of a population of Ypos chromosomes backcrossed from 129/Sv onto C57BL/6J genetic backgrounds for seven successive generations. Since all XYpos offspring were sterile, this line ended in the N7 generation. Traces Y females were absent in F1 but were present at high frequencies in litters from N2 and subsequent generations.

High Copy Y females were absent in N2 but appeared at increasing frequencies in the subsequent, N3–N7 generations. Ypos-linked repeated sequences that were stable during both germ-line and somatic development in the N4 and N5 generations underwent precise structural destabilization at high frequencies during somatic development in the N6 and N7 generations. Thus, maintenance of Ypos chromosomes on C57BL/6J backgrounds resulted in structural destabilization, which eventually inactivated functions of their $Tdy$ and other loci. The inactivation of testis-determining functions of the Ypos chromosome because of its sequence instability is predictably stochastic. Therefore, the mouse line being derived will end in sterile individuals after varying numbers of backcrosses of the Ypos chromosome onto the C57BL/6J genome. By selecting for males lacking any overt external signs of sex reversal, it has been possible to slightly extend the number of backcross generations to N10, before the line ends in sterile XYpos individuals (J. Barry Whitney III, personal communication). Others (Nagamine et al. 1987; Biddle and Nishioka 1988) have reported similar extensions of numbers of backcrosses of Ypos chromosomes onto C57BL/6J genetic backgrounds. However, the results presented in the present paper make it likely that C57BL/6J Ypos mice in each of these laboratories is genotypically distinct.

Stable X-linked transmission of Ypos-specific sequences from the germ line of a Low Copy Y female mated with males of an inbred strain was tested for. Nontransmission of Ypos-specific sequences from the Low Copy Y female may be explained by one of the following three observations: (1) There were 10-fold differences of copy numbers of Ypos-specific sequences in three tissues of this female. Their complete absence from ovaries may explain nontransmission of Ypos-specific sequences by this Low Copy Y female. (2) The results in the present paper showed that sequences on Ypos chromosomes underwent structural destabilization during germ-line and somatic development. Their excision from chromosomes of developing ova or embryos may explain nontransmission of Ypos-specific sequences from this Low Copy Y female. (3) Finally, other donor sequences or transgenes are also not transmitted from male (Wilkie and Palmiter 1987) and female (Rohan et al. 1990) germ lines of transgenic mice. In the case of the male mouse, the MyK 103 transgene had to be excised during
spermatogenesis (Wilkie and Palmiter 1987) to allow formation of viable sperm. Instability of other nucleotide sequences during somatic or embryonal development has also been reported (Kelly et al. 1989; Rohan et al. 1990).

Loss of Genetic Regulation of Recombination: Possible Cause of Instability of Y\textsuperscript{Pos} Chromosomes

Locus-specific and chromosome-specific genetic regulation of the five fundamental functions of eukaryotic chromosome biology—i.e., replication, repair, recombination (interchromosomal and intrachromosomal), condensation, and segregation—are well established in yeast, \textit{Neurospora}, and \textit{Drosophila} (Baker et al. 1976; Gatti et al. 1980; Strathern et al. 1981; Klapkoltz et al. 1985; Surosky and Tye 1988). Recombination in all organisms is known to be under rigorous genetic regulation (Kucherlapati and Smith 1988). The inbred strain C57BL/6J is genetically distinct from both 129/Sv and most other inbred strains (Taylor 1972; Cattanach and Moseley 1973; Forejt and Iványi 1975; Fitch and Atchley 1985). It is likely that alleles controlling recombination in C57BL/6J and 129/Sv inbred strains are distinct, thus explaining the instability of Y\textsuperscript{Pos} chromosomes on C57BL/6J but not on 129/Sv genetic backgrounds.

At least two distinct types of aberrant recombination events on Y\textsuperscript{Pos} chromosomes were identified by Y\textsuperscript{Pos}-specific sequences in Traces Y and High Copy Y females. Y\textsuperscript{Pos}-specific sequences in Traces Y females were first transmitted from F1 male germ lines. They were only partially representative of copies (~1/100) and types [A–C and (GATA)\textsubscript{5}] of interspersed repeated sequences from the entire length of Y\textsuperscript{Pos} chromosomes. Losses of flanking restriction sites of these traces of Y\textsuperscript{Pos}-specific sequences were also detected. These characteristics of Y\textsuperscript{Pos}-specific sequences are inconsistent with translocations, extrachromosomal elements, and recombinational hot spots. They are consistent with random interchromosomal recombination products which require limited sequence homology for heteroduplex formation, e.g., gene conversions. The inferred mechanism for the transfer of Y\textsuperscript{Pos}-specific sequences to XX females is the loss of suppression of X\textsuperscript{B6}-Y\textsuperscript{Pos} interchromosomal recombination. This inference is supported by other observations on X-Y recombination. X-Y interchromosomal pairing is known to be destabilized in offspring from matings between laboratory strains and wild mice (Cattanach and Moseley 1973; Forejt and Iványi 1975; Matsuda et al. 1983; Handel 1987). Many translocations from Y chromosomes are to X chromosomes (Singh and Jones 1982; Page et al. 1987; Petit et al. 1987), and unequal crossovers have been shown to occur in the murine pseudoautosomal region (Harbers et al. 1986).

Y\textsuperscript{Pos}-specific sequences in XY\textsuperscript{Pos} females showed contrasting characteristics. Precise reproducible losses or mobility shifts of all or most copies of multicopy Y\textsuperscript{Pos}-specific restriction fragments occurred in a single generation during somatic and/or germ-line development. These losses and mobility shifts were restricted to specific interspersed repeated sequences and to specific restriction fragments within a repeated sequence. Because of the high precision and frequency of these events and because of the large chromosomal sizes represented by interspersed repeated sequences, these were highly coordinated mechanisms. As somatic X-Y pairing is absent in mammals, these events must involve either intrachromosomal mechanisms of recombination or some other sequence modification.

The loss of control of recombination, which resulted in sequence instability, was not limited to combinations of the C57BL/6J inbred strain and feral Y\textsuperscript{Pos} chromosomes. Y chromosomes from the inbred strain C57BL/6J and another feral mouse,
Instability and Inactivation of Murine Y Chromosomes

M. domesticus (Centreville Light), were reciprocally backcrossed for two generations. Traces of Y-specific sequences and precise losses and mobility shifts of multicopy Y-linked sequences also resulted from both types of these matings (author’s unpublished data).

The high frequencies and precision of highly coordinated changes of sequences on Y sure chromosomes suggest that these changes were likely to result from mechanisms that were distinct from other mechanisms causing genetic instability. The loss of epigenetic control of retroviral replication by a maternal resistance factor and FV restriction loci in SWR/J × RF/J murine hybrids results in the induction of retroviruses (Jenkins and Copeland 1985). Similarly, their lower frequencies and different types of aberrant recombination products suggest that the meiotic instability of the polymorphic repeat sequence PR 1 and VrDNAs in the murine rDNA loci (Kuehn and Arnheim 1983; Kominami et al. 1985), microrecombinations in the murine major histocompatibility locus (Geliebter and Nathenson 1988), somatic instability of murine variable numbers of tandem repeats (VNTRs) (Kelly et al. 1989), and the rearrangement-induced premeiotically (RIP) in N. crassa (Selker et al. 1987) are mechanistically unrelated to the instability of the Y sure chromosome.

It is now recognized that unknown mechanisms result in two opposing effects on the repeated sequence fraction of the mammalian genome. Although there is a large extent of nucleotide sequence flux within it, structural and functional order are imposed on the mammalian genome (Goldman et al. 1984; Dover and Flavell 1984; Bernardi et al. 1985; Dover and Tautz 1986; Schmid and Shen 1986; Zuckerlandl 1986; Dover 1987, 1990; Korenberg and Rykowski 1988; Kucherlapati and Smith 1988). Fluxes of repeated sequences include such large genome-encompassing effects as their concerted evolution, retrotranspositioning (Dover and Flavell 1984; Baltimore 1985; Rogers 1985; Dover and Tautz 1986; Schmid and Shen 1986; Zuckerlandl 1986; Dover 1987, 1990), and expansions and contractions of VNTRs (Dover 1987, 1990; Jeffrey et al. 1990). The rate, precision, and extent of these nucleotide sequence fluxes cannot be explained either by Mendelian inheritance or by stochastic, slow, and unregulated processes such as drift, selection, and “selfishness” of DNA (Dover and Tautz 1986; Schmid and Shen 1986; Dover 1987, 1990). Loss of regulation of any of the unconventional mechanisms, either imposing structural and functional order on mammalian genomes or driving large nucleotide sequence fluxes through them, are potential mechanisms for the non-Mendelian transmission of Y sure-specific sequences documented in the present paper.

Loss of Y sure Chromosome Structure and Function as Consistent with Hybrid Dysgenesis Occurring in Mammals

When genomes of two evolutionarily diverged species are hybridized, a collection of genotypic and phenotypic dysfunctions are induced. Traits collectively characterizing hybrid dysgenesis are requirements of directionality of genetic crosses, induction of transposons, disruption of regulation of interchromosomal (meiotic) and intrachromosomal (mitotic) recombination, suppression of recombination at active loci, induction of recombination in inactive loci (and genomes), chromosome nondisjunction, segregation (or transmission-ratio) distortion, gonadal dysfunction (sterility), high frequencies of mutations, and embryonal lethalties (Haldane 1922; Dobzhansky 1951; Tracey 1972; Forejt and Iványi 1975; Kidwell et al. 1977; Bregliano et al. 1980; Bingham et al. 1982; Handel 1987; Yannopoulos et al. 1987). Most of the genotypic and phenotypic traits defining hybrid dysgenesis in other species are also identified in
the C57BL/6J Y\textsuperscript{Pos} mouse. Two of the most important traits characterizing hybrid dysgenesis are loss of control of recombination and testicular dysfunction, and they are specifically recognized in C57BL/6J Y\textsuperscript{Pos} mice (for a list of these and other traits, see table 2). Many of these traits were also reproduced when the Y chromosome from another feral species, \textit{M. domesticus} (Centreville Light), was backcrossed onto C57BL/6J inbred strains (author's unpublished results).

Like most inbred strains, C57BL/6J originated as a hybrid of \textit{M. musculus domesticus} ? \times \textit{M. m. musculus} \delta (Bishop et al. 1985). The C57BL/6J inbred-strain genome contains alleles that make it distinct from all other inbred strains (Taylor 1972; Cattanach and Moseley 1973; Forejt and Iványi 1975; Bonhomme et al. 1984; Fitch and Atchley 1985). The Y\textsuperscript{Pos} chromosome itself originates from \textit{M. poschiavinus} species, which is highly diverged from all laboratory strains, both in its karyotypic constitution (Gropp et al. 1972; Cattanach and Moseley 1973; Capanna 1982) and in its nuclear (Cattanach and Moseley 1973) and mitochondrial (Ferris et al. 1983) genomes. It is likely that (a) differences in C57BL/6J alleles controlling recombination and (b) the evolutionary divergence of Y\textsuperscript{Pos} chromosomes act together to produce sequence instability of the Y\textsuperscript{Pos} chromosome in this inbred strain.

\textbf{Table 2}

\textbf{Traits Characterizing Dysgenesis in Strains Derived from \textit{Mus musculus} × \textit{M. poschiavinus} Hybrids}

<table>
<thead>
<tr>
<th>Trait</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotypic:</td>
<td></td>
</tr>
<tr>
<td>Directionality of genetic crosses: C57BL/6J; ? × 129/Sv Y\textsuperscript{Pos} ; \delta causes dysgenesis</td>
<td>Present study</td>
</tr>
<tr>
<td>Induction of interchromosomal recombination: Y-linked sequences are transferred to other chromosomes (transferred sequences are derived from B1, R, and satellite repeats) and sequences from &gt;6,000 kb are transferred from Y\textsuperscript{Pos} chromosome</td>
<td>Nallaseth 1987; present study</td>
</tr>
<tr>
<td>Suppression of interchromosomal recombination in autosomal and X chromosomal loci</td>
<td>Cattanach and Moseley 1973</td>
</tr>
<tr>
<td>Aneuploidy of autosomes and X chromosomes</td>
<td>Gropp et al. 1972; Cattanach and Moseley 1973; Capanna 1982; Magnuson et al. 1985</td>
</tr>
<tr>
<td>Increased frequency of transmission of 4.7-kb EcoRI autosomal/X chromosomal unique sequence</td>
<td>Nallaseth 1987</td>
</tr>
<tr>
<td>Induction of mitotic instability of Y\textsuperscript{Pos} chromosomes</td>
<td>Nallaseth 1987</td>
</tr>
<tr>
<td>High frequency of mutations</td>
<td>Cattanach and Moseley 1973</td>
</tr>
<tr>
<td>Phenotypic:</td>
<td></td>
</tr>
<tr>
<td>Disruption of primary sex determination</td>
<td>Eicher et al. 1982; present study</td>
</tr>
<tr>
<td>Disruption of oogenesis and spermatogenesis</td>
<td>Gropp et al. 1972; Cattanach and Moseley 1973; Capanna 1982; Eicher et al. 1982; Magnuson et al. 1985; Nallaseth 1987</td>
</tr>
<tr>
<td>Disruption of embryogenesis</td>
<td>Gropp et al. 1972; Cattanach and Moseley 1973; Magnuson et al. 1985</td>
</tr>
<tr>
<td>Disruption of secondary masculinization</td>
<td>Eicher et al. 1982; Nallaseth 1987</td>
</tr>
</tbody>
</table>
Two important differences are noted between the hybrid dysgenesis detected in C57BL/6J Y\textsuperscript{Pos} mice and that in Drosophila (Kidwell et al. 1977; Bregliano et al. 1980; Bingham et al. 1982; Yannopoulos et al. 1987). First, transposon (or retroviral) excision does not explain the presence of Y\textsuperscript{Pos}-specific sequences in females. Second, in addition to random recombination, precise, reproducible, and high-frequency sequence instabilities are induced de novo in loci that are normally recombinationally suppressed on the Y chromosome during hybrid dysgenesis. Only imprecise recombination at chromosome breaks marking sites of transposon excision occurs during hybrid dysgenesis in Drosophila. Finally, since natural hybrid zones of M.\textit{musculus} × M.\textit{poschiavinus} have been identified (Gropp et al. 1972; Capanna 1982), these observations may have important implications for speciation both in these hybrids and in laboratory strains, which have originated as hybrids (Bishop et al. 1985).

Acknowledgments

I wish to thank J. Barry Whitney III, who initially provided 129/v Y\textsuperscript{Pos} and C57BL/6J Y\textsuperscript{Pos} male mice that were later transferred to the colonies at the University of South Carolina, and Michael J. Potter for providing \textit{Mus poschiavinus} (Valende) and \textit{M. domesticus} (Centreville Light) mice. They were maintained by funds from NCI, contract N01-CB-25594. I also gratefully acknowledge the many contributions of Peter J. Hornsby, Mel DePamphilis, Gerald Smith, Robert Erickson, Bruce Baker, Rollie Harp, Gloria Choice, Charlotte Joyner, and Sharon Perry that made this work possible. This work was partially supported by Public Health Service grant HD 17523 awarded to Michael J. Dewey.

LITERATURE CITED


Instability and Inactivation of Murine Y Chromosomes


WALTER M. FITCH, reviewing editor

Received February 6, 1991; revision received October 3, 1991.

Accepted October 4, 1991
Letter to the Editor

The Number of Replications Needed for Accurate Estimation of the Bootstrap P Value in Phylogenetic Studies¹

S. Blair Hedges
Department of Biology and Institute of Molecular Evolutionary Genetics, Pennsylvania State University

The bootstrap is a statistical method for obtaining a nonparametric estimate of error (Efron 1979, 1982). Felsenstein (1985) was the first to apply this method to phylogeny estimation, and his approach is now widely used. Taxa are held constant, and the characters (for sequence data, nucleotide or amino acid sites) are resampled randomly with replacement. A phylogeny is constructed from each replication of the data, and the frequency of appearance of particular phylogenetic groups (groups of alleles or taxa) among all of the trees constructed by this resampling is the bootstrap confidence limit, or bootstrap P value (BP). The BPs of different nodes within a tree can be used to assess the relative stability of those phylogenetic groups or, if applied strictly, to test their statistical significance (e.g., at the 95% or 99% level). The application of bootstrapping to phylogeny estimation is a tradeoff between the maximum number of replications that can be performed by the researcher in a reasonable amount of time and the minimum number of replications needed for accurate estimation of the BP. The purpose of the present report is to explore the variance (and hence the accuracy) of the phylogenetic BP and to establish guidelines for efficient bootstrap sampling.

BP is the proportion of trees containing a particular phylogenetic group. It therefore follows the binomial distribution, which has a variance of $\sigma^2 = [P(1 - P)/n]$, where $P$ is the BP and $n$ is the number of replications. Although Li and Gouy (1990) recently suggested that more replications are needed for larger numbers of taxa, the accuracy of the BP is a function only of $P$ and $n$. If the interval containing 95% of the samples (±1.96 standard deviations) is used as a measure of accuracy, then the application of the above formula shows that 1,825 replications ($=0.95 \times 0.05(1.96/0.01)^2$) are needed to attain ±1% accuracy at a BP of 0.95 (fig. 1). This is more than an order of magnitude higher than the number of replications (50-100) normally used in phylogenetic analyses.

Based on this, a practical guideline for efficient and accurate bootstrap sampling can be made: If one wishes the expectation to be that the 95% confidence range is ±1% of the BP, then one must perform 2,000 bootstrap replications (if BP = 0.95) or 400 replications (if BP = 0.99) in phylogenetic analyses, unless the computational time is prohibitive; additional replications are unnecessary, and fewer replications may sacrifice statistical accuracy. Moreover, statistical testing at the 95% level cannot be made using <73 replications, even if the group is supported by a BP of 1.00. This is because the inaccuracy at a mean BP of 0.95 is greater than ±5% (fig. 1) when <73 replications are used. In other words, a BP ≥1.00 could be obtained when the actual (mean) BP is not significant (<0.95). Thus, the 50 replications used by Felsenstein (1985) in his original example and the 20–100 replications used in many subsequent studies (e.g., see Ovenden et al. 1987; Thomas et al. 1989; Jansen et al. 1990; Meyer and Wilson 1990; Douglas et al. 1991; Irwin et al. 1991) would appear to be far too few for the intended purpose of statistical testing at the 95% level.

1. Key words: phylogeny, statistics, DNA sequence, systematics, evolution.

Address for correspondence and reprints: S. Blair Hedges, Department of Biology, 208 Mueller Lab, Pennsylvania State University, University Park, Pennsylvania 16802.

© 1992 by The University of Chicago. All rights reserved.
0737-4038/92/0902-0013$02.00
366
FIG. 1.—Accuracy of bootstrap P value (±1%, ±2%, and ±5%) vs. number of bootstrap replications, based on binomial variance. The dashed line is the 0.95 P value (95% confidence limit). Top, Region spanning P values 0.5–1.0. Bottom, More detailed plot spanning P values 0.90–1.00.
The number of bootstrap replications needed was addressed recently by Efron (1987). He showed that, in the case of bootstrap confidence intervals, the coefficient of variation is substantial (9%) with only 200 replications, whereas it decreases to 4% with 1,000 replications (the number of replications needed for calculating the bootstrap standard error is considerably fewer, \( \sim 100 \), but this statistic is of limited value in phylogeny estimation). Although Efron (1987) recommended that "on the order of 1000" replications are needed, it has been shown here that the actual phylogenetic BP may be over- or underestimated by 1%-2% in the region of the 0.95 BP with 1,000 replications. Some would consider this an acceptable error, but it would mean that the researcher would be unable to state (validly) that a group supported by a 0.96-0.97 BP is statistically significant. With 2,000 replications, such a statement can be made.

The typical size of data sets used in phylogeny estimation will almost certainly increase as more sequences become available. This may place a computational constraint on the number of bootstrap replications possible in large data sets. However, the bootstrap method still can be used even when only a small number of replications is feasible, as long as the variance of the BP is taken into consideration when one is drawing conclusions. BPs with \( \pm 6\% \) accuracy can be obtained with only 50 replications (in the 0.95 region), and, although this error is too high for statistical testing, it can provide a reasonable indication of relative stability of groups within a phylogenetic tree, especially if no other statistical methods are available.

Only one aspect of the bootstrap method has been considered here: the number of replications necessary for accuracy. Other limitations of this method must be considered in any application. As noted by Felsenstein (1985), a substantial lack of independence of characters within the data set may require an adjustment in the sampling method, such as sampling fewer than the total number of characters randomly, with replacement, from the total number of characters. Also, the BP associated with a node reflects only that particular data set and clustering method. For example, either high levels (e.g., \( >50\% \)) of sequence divergence or considerable rate variability among lineages may lead to statistical inconsistency with most methods of tree construction (Felsenstein 1983, 1988; Li and Gouy 1990). In these cases, bootstrapping could show statistically significant support for an incorrect topology (Nei 1991). If these limitations are kept in mind, the bootstrap method can be a simple and effective means of evaluating the results of phylogenetic analysis.

Acknowledgments

I thank Linda R. Maxson (L.R.M.) and Masatoshi Nei for use of research facilities; Sudhir Kumar, Tatsuya Ota, and an anonymous reviewer for many helpful suggestions; and especially Masatoshi Nei for invaluable statistical advice. This work was supported by National Science Foundation grant BSR 8918926 to L.R.M. and S.B.H.

LITERATURE CITED


BRIAN CHARLESWORTH, reviewing editor

Received May 3, 1991; revision received July 12, 1991

Accepted September 9, 1991
Letter to the Editor

Substitution Bias, Weighted Parsimony, and Amniote Phylogeny as Inferred from 18S rRNA Sequences

Charles R. Marshall
Department of Earth and Space Sciences, University of California

Recently published rRNA sequences (Hedges et al. 1990) provide important new information on tetrapod relationships. Specifically Hedges et al.'s (1990) data support a monophyletic Lissamphibia, Amniota, Squamata, Mammalia, and Aves. It is surprising that the 18S rRNA sequences also support a close relationship between the two birds and the four mammals sequenced. Paleontological data provide considerable evidence for a close bird-crocodile relationship (e.g., see Gauthier et al. 1988; Donoghue et al. 1989), and the strength of these data suggested the possibility that the support for a bird-mammal relationship shown by the 18S rRNA sequences may be an artifact created by substitution biases.

Table 1 shows that there are indeed pronounced substitution biases in the amniote 18S rRNA sequences. In particular, there is both a significant overrepresentation of T-to-C substitutions and a significant underrepresentation of A-to-T and T-to-A substitutions. It is interesting that a disproportionately large number of sites that support a bird-mammal relationship are T-to-C substitutions (table 1). The only site that involves T-to-A and/or A-to-T substitutions is site 1317 in Hedges et al.'s (1990) figure A1 (note that there are 140 nucleotides/line in their fig. A1, not 150 as indicated), and on the most parsimonious tree this site is homoplasious, requiring both a T-to-A substitution in the branch leading to the four mammals and either a T-to-A substitution in the Coelacanth or an A-to-T substitution at the base of the amniotes. However, when viewed in isolation, site 1317 supports a clade including birds and crocodiles, to the exclusion of mammals.

Note that the substitution biases are quite complex. The T/C transitions are vastly overrepresented, yet A/G transitions are not. Similarly, A/T transversions are underrepresented, but C/G transversions are not (table 1).

Dynamically Weighted Parsimony

Williams and Fitch's (1990) dynamically weighted parsimony method provides a way of determining whether the substitution biases in the 18S rRNA sequences might be responsible for the unexpected 18S rRNA amniote phylogeny. Their algorithm weights sites according to the relative number of substitutions each incurs on the shortest tree, and it weights substitution types according to number of times they appear on the shortest tree; a site involving relatively few substitutions and relatively rare substitution types will be assigned a higher weight than will a site requiring many substitutions involving frequently seen substitution types. Of course, both the number of substitutions that a site incurs and the frequency of each substitution type depend in part on the topology of the tree being examined. The algorithm assigns weights on the basis of a predefined topology and then uses these weights to search for a new most parsimonious tree. If a new topology is found, then the weights are recalculated, and a new shortest tree is searched for, etc, i.e., the weights are assigned dynamically.

1. Key words: 18S rRNA, amniote phylogeny, birds, crocodiles, mammals, weighted parsimony, substitution bias.

Address for correspondence and reprints: Charles R. Marshall, Department of Earth and Space Sciences, University of California, Los Angeles, California 90024-1567.
### Table 1
Numbers of Substitutions on Most Parsimonious Tree for 18s rRNA Amniote Sequences

<table>
<thead>
<tr>
<th>Substitution Type</th>
<th>All Amniote Branches</th>
<th>Birds and Mammals</th>
<th>Four Mammal Species</th>
<th>No. of Character Changes on Entire Tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-C</td>
<td>20***</td>
<td>6***</td>
<td>7***</td>
<td>59***</td>
</tr>
<tr>
<td>C-T</td>
<td>10**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A-G</td>
<td>6</td>
<td>1</td>
<td>4</td>
<td>24</td>
</tr>
<tr>
<td>G-A</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>A-T</td>
<td>0**</td>
<td>0</td>
<td>0</td>
<td>2***</td>
</tr>
<tr>
<td>T-A</td>
<td>0**</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>A-C</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>C-A</td>
<td>1*</td>
<td>0</td>
<td>0</td>
<td>12*</td>
</tr>
<tr>
<td>T-G</td>
<td>5</td>
<td>0</td>
<td>3</td>
<td>11**</td>
</tr>
<tr>
<td>G-T</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>G-C</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>17</td>
</tr>
<tr>
<td>C-G</td>
<td>7</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Expected no. of changes/substitution type</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>21</td>
</tr>
</tbody>
</table>

**NOTE.**—For over-/underrepresented substitution types the probability that the observed number or a greater/fewer number of substitutions should be observed was calculated using the binomial distribution. Significant deviations from the expected number of substitutions per substitution type are indicated by the asterisk(s) (see P-value footnotes below).

* Derived condition is in the second position and in boldface type.

b Substitution types are grouped so that unpolarized substitutions on the branch leading to Coelacanth could be included.

* Rounded to nearest integer and calculated under the assumption that all substitution types are equally likely to occur at each site.

* * * P < 0.05.

** P < 0.005.

*** P < 0.0001.

As suspected, when the amniote 18s rRNA sequences were analyzed with Williams and Fitch's weighted parsimony program (WTSUBS), the shortest tree differed from the tree produced by unweighted parsimony and is almost identical to the "paleontological" tree, with just the birds and squamates switched in position (fig. 1). The weighted parsimony tree (fig. 1) was found by WTSUBS in all 21 of its different combinations of parameters, when given the "paleontological" tree as the starting tree. WTSUBS became trapped in a local minimum when given the most parsimonious tree (unweighted) as the starting topology and did not find a shorter tree even though this starting topology is ~6% longer than the weighted parsimony tree shown in figure 1.

The difference between the dynamically weighted parsimony tree and the paleontological tree is troubling, but the discrepancy may simply be due to the small number of informative sites available for the Amniota. At site 1317 (discussed above) the sequence is only known for the four mammals (A), for one of the birds (T), for the crocodile (T), and for the outgroup (A). If the lizard, snake, and turtle are found to have the outgroup condition (A) at this site, then weighted parsimony may well unite the birds and crocodile as closest relatives. However, the fact that the topology of the shortest tree may be sensitive to the sequences at just one site suggests that the 18s rRNA molecule is really too highly conserved to be particularly informative when one is trying to assess the relationships between the amniote classes.
"Paleontological" phylogeny

Coelacanth
Mammals
Turtle
Squamates
Crocodilian
Birds

18S rRNA maximum parsimony phylogenies

Unweighted
Coelacanth
Turtle
Squamates
Mammals
Birds

Weighted
Coelacanth
Mammals
Turtle
Squamates
Crocodilian
Birds

FIG. 1.—Alternative phylogenies of major amniote classes. The “paleontological” tree is based on a maximum-parsimony analysis of morphological characters drawn from fossil and extant taxa (Gauthier et al. 1988). The unweighted parsimony tree is paleontologically highly implausible, whereas the weighted parsimony tree is almost identical to the “paleontological” tree. The major controversy in amniote relationships centers around the groups shown in boldface type. The 18S rRNA sequences analyzed were from *Latimeria chalumnae* (coelacanth), *Pseudemys scripta* (turtle), the two squamates *Sceloporus undulatus* (lizard) and *Heterodon platyrhinos* (snake), *Alligator mississippiensis* (crocodilian), the two birds *Gallus gallus* (galliform) and *Turdus migratorius* (passeriform), and the four mammals *Oryctolagus cuniculus* (rabbit), *Rattus norvegicus* (rat), *Mus musculus* (mouse), and *Homo sapiens* (human).

Bootstrapping in the Light of Substitution Biases

Bootstrapping allows one to gauge the relative strengths of mutually contradictory phylogenetic signals in a data set. However, if the data are biased, then the bootstrapping support may also be biased and should not be interpreted as a statistical measure of confidence in the hypothesis being evaluated. In light of the nonrandomness in the amniote 18S rRNA substitutions, the fact that 88% of bootstrap maximum-parsimony trees show support for a bird-mammal association (Hedges et al. 1990) may just as likely reflect the strength of the systematic biases in the amniote sequences as reflect the likelihood that birds are the closest relatives of mammals.

Acknowledgments

Thanks to Dr. Hedges for providing the aligned sequences; to Norman Pace, Clint Turbeville, and Rudy Raff for helpful criticisms; to Walter Fitch and Eileen Owens for running WTSUBS; to Walter Fitch for making numerous suggestions for improving the manuscript; and to Rudy Raff and the Department of Biology, Indiana University, for providing the facilities for preparing the manuscript. This research was supported by NIH NRSA fellowship 5 F32 HD07332-02.

LITERATURE CITED


WALTER M. FITCH, reviewing editor.

Received February 14, 1991; revision received June 20, 1991.

Accepted June 28, 1991
18S rRNA Sequences and Amniote Phylogeny: Reply to Marshall

S. Blair Hedges and Linda R. Maxson
Department of Biology and Institute of Molecular Evolutionary Genetics, Pennsylvania State University

Marshall (1992) presents a reanalysis of some 18S rRNA data (Hedges et al. 1990) bearing on tetrapod relationships. Our phylogenetic analyses of these data for 26 tetrapods, supported among other groupings, a bird-mammal relationship. However, Marshall concluded that his reanalysis of the amniote data, using a weighted parsimony method (Williams and Fitch 1989, 1990), supported a tree uniting birds and squamates (lizards and snakes). We believe that his results were due to a misapplication of the weighted parsimony method, which was designed for highly variable (noisy) data—not for highly conserved data.

The controversy over amniote relationships involves not only the palaeontological data and our 18S rRNA data but other morphological and molecular data sets (Hedges et al. 1990; Hedges and Maxson 1991). The best estimate of amniote phylogeny will not hinge on any one data set unless such a set is very large; more sequence data will be needed to resolve this important question. However, the available 18S rRNA sequence data (Hedges et al. 1990, fig. A1; present paper, fig. 1) clearly support a bird-mammal relationship.

Weighted parsimony methods were developed by Sankoff and Cedergren (1983) and Williams and Fitch (1989, 1990), to account for substitution biases in phylogenetic analysis. Marshall (1992) used the latter method to reanalyze our 18S rRNA data. The basic principle is that rare substitution types should be weighted more heavily than common substitution types because common substitutions are more likely to occur multiple times at the same site, thus obscuring phylogenetic information. The implementation of the weighting is done a posteriori by observing the frequency of the different substitution types in an initial tree and then weighting inversely to those frequencies. Each site also can be weighted inversely to the number of changes at that site in the initial topology.

That there is substitution bias in the 18S rRNA data is not surprising, because it is present in most nucleotide sequence data sets. Mechanisms have been proposed for some types of biases (e.g., transition-transversion bias and codon bias), but the reason for the unequal frequencies of certain substitution types in the 18S rRNA data is presently unknown.

One potential problem with weighted parsimony involves the basic assumption that rare substitution types are more reliable indicators of phylogenetic relationships. This concept is more useful when there is a high probability of multiple changes per site (multiple hits). Weighting rare changes more heavily in a highly conserved set of sequences (such as the 18S rRNA data) is unwarranted because all substitution types convey the same phylogenetic information (i.e., they are equally detectable) regardless of relative frequency. There is a wide "gray zone" where the relative information content of rare changes increases as the probability of multiple hits increases. It is unclear at what point (if any) an inverse weighing scheme proportionately compensates for the increasing noise in the data. Application of such a weighting scheme to a data set which has not reached this noise level constitutes a bias.

Williams and Fitch (1989, 1990) intended their method to be used with noisy data sets. Although the presence of homoplasy in the 18S rRNA data indicates that...
some multiple changes have occurred, the low level of sequence divergence (4.4% between amphibians and amniotes) and the low level of three-variant (1.0%) and four-variant (0.2%) sites across 26 taxa indicates that the data set is not noisy. Therefore, the weighting scheme imposed by Marshall (1992) is unwarranted.

Another problem with weighted parsimony, recognized by Williams and Fitch (1989, 1990), is seed-tree topology bias: the output tree depends, to a varying degree, on the input tree. Williams and Fitch (1989) recommended using several different seed trees if one is unsure of the true phylogeny. Marshall used only two seed trees. With the "paleontological" seed tree (birds + crocodilians) he obtained a tree joining birds and squamates, and with the unweighted seed tree (birds + mammals) he obtained the seed topology. Because the birds + squamates tree was 6% shorter than the birds + mammals tree, Marshall concluded that weighted parsimony supports a bird-squamate relationship.

We interpret these results differently. Tree length in a weighted parsimony analysis does not have the same meaning as tree length in a conventional (unweighted) parsimony analysis. It is not a measure of the actual number of substitutions in a tree; rather, it is a compounded value that depends on relative frequencies both of substitution types and of changes per site. It is debatable whether the two trees obtained by
Marshall, each using a different seed tree, can be validly compared. We believe that they cannot be compared, because of the influence of the seed-tree topology on tree length. Even if one argues that they are comparable, it must be shown that the two trees are significantly different in length. Given the results (two input trees, two output trees), one wonders whether every different input tree will result in a different output tree—and whether some third seed-tree topology might result in yet a shorter output tree!

Because of the recognized seed-tree topology bias of weighted parsimony, we believe that the only seed tree that is justifiable is the initial unweighted tree, which in this case is the bird + mammal tree. In that no better tree could be found by using that tree as the seed tree, we interpret Marshall's results as affirming that the 18S rRNA data support a bird-mammal relationship. However, other problems—including the use of the coelacanth, rather than the closer lineage (amphibians), as the outgroup—preclude any interpretations from Marshall's reanalysis.

The influence of the “paleontological” seed-tree topology on the results obtained by Marshall underscores how previous hypotheses of relationships can effect phylogenetic analysis. Our analyses of amniote relationships are unbiased, and we have found support for either mammals (Hedges et al. 1990) or crocodilians (Hedges and Maxson 1991) as the sister group to birds. If future studies show overwhelming support for a bird + crocodilian relationship, then it will be interesting to determine why and how the bird and mammal sequences of several genes (beta hemoglobin, myoglobin, and 18S rRNA) have converged. On the other hand, if a bird + mammal relationship is later confirmed, then it will be interesting to determine how the bird and crocodilian sequences (histone H2B and pancreatic polypeptide) have converged—and how the fossil record can be reinterpreted. Whatever final consensus is obtained, phylogenetic analysis should be independent of previous hypotheses.

Despite our criticisms of weighted parsimony as applied by Marshall, we recognize the importance of understanding and accounting for the biases inherent in sequence data. At present, the biases and the mechanisms responsible for those biases are not well understood. Methods of sequence analysis will surely improve when those mechanisms become better known.

Acknowledgments

This research was supported by a grant from the National Science Foundation (BSR 8918926 to L.R.M. and S.B.H.).

LITERATURE CITED


WALTER M. FITCH, reviewing editor

Received August 28, 1991; revision received September 3, 1991
Accepted September 3, 1991
Commentary on Letters of Marshall and of Hedges and Maxson

Hedges and Maxson contend that weighting should only be performed if the data are noisy, i.e., homoplasious. I believe weighting is, in principle, appropriate for all data and offer the following thought experiment. Suppose we have 24 mitochondrial sequences, α–ω, 10 kb long, from which we are able to extract 527 cladistically informative positions. They give a tree for which there are 20 character-state changes supporting every clade except one, and bootstrapping supports each of these clades at >99%, except for that clade which involves sequences α, β, and γ. There are only two positions, call them 1 and 2, that address the issue of the unresolved αβγ clade. They are shown in figure 1, along with a transformation matrix that shows the distribution of the nucleotide substitutions in the other 525 positions. (In the table the numbers s/v are the number of transition changes/the number of transversion changes going from the nucleotide on the left of the row to the nucleotide at the top of the column.) Note from the matrix that (1) transitions outnumber transversions 4 to 1 (420 to 105) and (2) there are no homoplasies among the transversions, (Y = R), but that 20% of the transitions, (A = G) and (C = T), are parallel to another 20% of them. Characters 1 and 2 contradict each other. There are two possible resolutions, equally parsimonious under uniform weighting, for the αβγ clade shown in the figure. Are they equally likely? It is a choice between a tree requiring parallel T → C substitutions, which are very common, or a tree requiring parallel T → A substitutions, which are most uncommon. The former seems more probable. Weighting does nothing more than attempt to increase the odds that you select the correct tree, by assigning weights reflective of the properties of the very data being analyzed. If that consideration is apt, then weighting is generally apt. This is not to say we know how best to weight. This is not to say the expected difference will always be in the correct direction, but it should, on average, be an improvement. It may even make some differences significant that would not otherwise have been. The improvement may not always be worth the added effort, but computers make the added effort very small. If weighting is generally apt, then how noisy the data are is irrelevant.

Hedges and Maxson assert that the seed-tree topology influences the final tree’s length. This is untrue. The seed tree is used not for initial weighting but only as the starting point for searching for the best tree it can find. This is the first pass. The best tree found is then used to assign weights for a subsequent pass. In the first pass, weights, if not uniform, are computed on the basis solely of nucleotide frequencies and are thus independent of tree topology. The last pass is last because it obtained the same best tree as did the previous pass. This means that its length is determined by weights assigned on the basis of an optimum fit of the data to the tree. If one has two different

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig1.png}
\caption{Fig. 1.}
\end{figure}
“best” trees because of different seed trees, it is simply because each was in a local optimum from which it could not escape. But the score for each best tree, being based on weights optimized for its own topology, can be legitimately compared. Weighted parsimony asks, Which tree, when it itself is used as the basis for assigning weights, is most parsimonious? There may be a tree better than that proposed by Marshall, but it remains true, on the basis of the rules used to assign weights, that his tree was preferred to the birds plus mammals tree. That does not prove Marshall is correct, but it does mean that the difference between the authors, as expressed in their letters, seems to reside solely in their readiness to weight characters and their transformations.
Information for Contributors

Molecular Biology and Evolution is a bimonthly journal devoted to the interdisciplinary science between molecular biology and evolutionary biology. The journal will emphasize experimental papers, but theoretical papers are also published if they have a solid biological basis. Although this journal is primarily for original papers, review articles and book reviews are also published. Brief discussion and comment on material published in this journal or on issues particularly relevant to readers of this journal will be published as "Letters to the Editor." Letters that refer to a paper handled by an associate editor should be sent to that editor or to the editor in chief.

To minimize publication delays, authors should follow the instructions given here and should also provide their telephone numbers.

Submission of Manuscripts

Send manuscripts (one original and two high-quality copies) to the editor in chief, managing editor, or any associate editor. Any manuscript or any part of a manuscript that has been published or submitted for publication elsewhere cannot be accepted for publication. Anyone who wishes to write a review article should contact the managing editor. Correspondence about book reviews should also be addressed to the managing editor. Decision on acceptance of papers will be made as rapidly as possible. Papers that are not suitable to the journal will be returned immediately to authors without detailed review.

After a manuscript is accepted, its author will be requested to sign an agreement transferring copyright to the publisher. No published material may be reproduced or published elsewhere without the written permission of the copyright owner. The journal will not be responsible for the loss of manuscripts at any time.

Publication is taken to imply that the authors are prepared to make available to the public any clone of cells, DNA, or antibodies used in reported experiments and any computer programs or nucleic acid or protein sequences on which the paper is based. Nucleic acid sequences must be provided to GenBank in computer-readable format along with the appropriate associated data. The GenBank accession number will be printed in the article.

Preparation of Manuscripts

Papers must be written in English and organized in the sequence described below. Each section must be typed double-spaced on heavyweight, nonerasable bond; the page margins should be 1½ inches wide to allow for corrections, manuscript editor’s notes and queries, and typesetting markup. Type should not be more than 12 characters per inch nor more than six lines per inch. Special typefaces (e.g., italic or sans serif) should not be used, and right-hand margins should not be justified. Word-processing output on dot matrix printers is acceptable only if it is the quality of the standard typewriter. Handwritten items (e.g., Greek letters) must be identified in the margin. Non-English words must have correct diacritics. Although each major part of the paper (e.g., Literature Cited) must begin on a new page, the pages should be numbered consecutively throughout, beginning with the title page and continuing through the abstract, text, Appendix, Literature Cited, footnotes, tables, and ending with figure legends.

Title page.—This page should contain the paper’s title, the names of all authors, the institution(s) at which the research was done, the current affiliations of all authors, the name and address for correspondence, and a footnote on nonstandard abbreviations used, if any (see below). Finally, the title page should also provide a running head (maximum of 50 characters and spaces).

Abstract.—The abstract should be a one-manuscript-page factual condensation of the entire paper, including a statement of purpose, a clear description of observations and findings, and a concise presentation of conclusions. It should not assert that the findings are discussed.

Key words.—A list of three to six words or phrases should be provided that will accurately index the subject matter of the article.
Text.—The text should comprise the following sections: (1) Introduction, (2) Material and Methods, (3) Results, (4) Discussion, and (5) Acknowledgments (if any). Papers should be concise but will not be restricted in length.

All organisms mentioned must be identified by their scientific binomens and underlined. Symbols for genetic loci must also be underlined and should follow the established rules of genetic nomenclature for the various organisms (consult M. Demerec et al., *Genetics* 54:61–76, 1966). Include the formal IUB name and number of all enzymes mentioned.

Do not use abbreviations for words or phrases used less than five times. Abbreviations used by the *Journal of Biological Chemistry* will be regarded as standard; nonstandard abbreviations should be defined collectively in a footnote.

Mathematical equations must be carefully typewritten: spacing between characters should be correct as typed. It will be assumed that all characters in equations and their counterparts in the text will be set in italics unless the author specifies otherwise the first time a character appears. Equations should be numbered sequentially, in arabic numerals in parentheses, on the right-hand side of the page.

These and other guidelines can be found in the *Council of Biology Editors Style Manual* (4th ed., 1978). In general, all material should conform to the CBE format. See also recent issues of this journal.

Terminology.—A satisfactory interdisciplinary communication requires using words with careful attention to their precise meaning in both disciplines. Authors may use any word they choose provided only that its meaning is clear, consistent, and serves to increase the paper's comprehensibility. The following preferred usages will be assumed unless authors define them otherwise.

Where the alignments disagree, they are differences rather than changes, since there may have been multiple changes to create a single difference. Differences or changes are replacements if the sequences are amino acids, substitutions if they are nucleotides. The word “invariant” has two common but different meanings, invariable and unvaried. Choose one of the latter two. Mutations should be restricted to changes before selection has operated. If two molecules are alike in some degree, they are similar. If you infer from their similarity that they have a common ancestor, then they are homologous, but, if their similarity was acquired by convergence, they are analogous. When homology arises via a gene duplication (all or part), it is paralogy; when it arises via speciation, it is orthology; when it arises by horizontal gene transfer, it is xenology. The phrase “insertions and/or deletions” may be reduced to indels. Gaps are introduced into sequences to increase their similarity rather than to optimize similarity, unless an algorithm is employed that guarantees an optimized result according to the way similarity is defined (e.g., as maximum matches). Similarity should not be asserted to be significant unless patently obvious or accompanied by a probability statement and its method of determination ($\chi^2$, standard measure, binomial, etc.).

Trees determined by methods that can produce branch lengths should have those lengths placed upon them. Bootstrapping is encouraged, and those numbers should be placed on the
tree also (see example). An alternative to bootstrapping is the provision on the tree of the estimated variance of each leg's length. If more than one tree is obtained for the same taxa, rotations about ancestral nodes should be done so as to make the order of the taxa agree as much as possible with each other. If there is a table of distances or a set of aligned sequences, the order of the taxa should also match that of one of the trees obtained from them.

The IUB single-letter code for nucleotide bases including ambiguity is: A = adenine, C = cytosine, G = guanine, T = thymine, U = uracil, R = A/G (purine), Y = C/T (pyrimidine), M = A/C, W = A/T, S = C/G, K = G/T, B = C/G/T (not A), D = A/G/T (not C), H = A/C/T (not G), V = A/C/G (not T), N = X = A/C/G/T (any or unknown). For ambiguous nucleotides, T and U are equivalent.

**Literature Cited.**—Literature in the text should be cited by author and year and, where citation is to a book, the relevant pages thereof. Text citations of two or more works at a time should be given in chronological order. When a paper written by three or more authors is cited, write the name of the first author plus “et al.” The Literature Cited section at the end of the paper should be arranged alphabetically by author(s) and then chronologically and should contain only works specifically cited in the text. References to papers that have not yet been published will be as for articles (see below), except that “submitted” or “accepted” (along with the journal name) will replace the volume and page numbers. “In press” will not be used. (When such papers include authors of the submitted manuscript, copies of those papers must accompany the submitted manuscript.) Material that has not been submitted should not be referenced in the Literature Cited section but, rather, should simply be given the notation “unpublished data” at the appropriate place in the text.

For the style of citations, please note the following examples:

**Journal articles:**


**Books:**

**INGRAM, V. M. 1963. The hemoglobins in genetics and evolution. Columbia University Press, New York. (The pages of the book that are appropriate to the item cited must be given in the text where the book is cited.)**

**Book chapters:**


The abbreviations of periodicals should be those used in *Chemical Abstracts*. Articles should include “Reviewed under the auspices of (name of the editor or associate editor with whom author has corresponded)” at the end of the Literature Cited section.

**Footnotes.**—Footnotes should be used sparingly. When necessary, they should be indicated in the text by superscript arabic numerals; the notes themselves should be typed on a page separate from the text. Footnotes to tables are referenced by superscript letters, except for significance levels, which use asterisks; table footnotes should be typed on the same page as the table to which they pertain.

**Tables.**—Each table must have a brief and self-explanatory title, be numbered with arabic numerals in order of its appearance in the text, and be typed on a separate page. Large, complex tables are discouraged. Guidelines for table format may be found in the *CBE Style Manual* and the *Chicago Manual of Style* or may be obtained by writing the editors of this journal.

**Legends.**—Figure legends should be typed on pages at the end of the manuscript, after tables. Each legend must be descriptive so that the illustration can be understood apart from the text and must define abbreviations used in the illustration.

**Illustrations.**—Each illustration (figure) should be an original, not a photocopy. Illustrations should be separate and have uniform lettering. They should be numbered consecutively, following the sequence in which they are mentioned in the text. The place where each illustration is to be inserted may be indicated by a circled note in the margin of the typescript. Names of authors, figure number, and an arrow indicating proper orientation should be written lightly in pencil on the back of each figure. Line drawings must be of high quality; typewritten or
hand lettering is unacceptable. Photographs should be high-contrast, glossy prints. Magnifications may be indicated by a micron bar or in the legend. Photographs to be reproduced without further reduction must be so marked and may not exceed 5 inches wide by 8 inches long (122 by 194 mm), in order to fit the journal format. Please keep in mind that these dimensions are maxima. Sequence alignments must not exceed two pages when published. Dot matrices and hydropathy plots are rarely accepted. Colored illustrations are acceptable but will incur a moderate charge.

Proofs and Reprints

Proofs and reprints will be sent to the author (or, in the case of multiple authors, to the senior author) with page proofs.

Page Charges

There is a charge of $25 per page or fraction of a page for publication in Molecular Biology and Evolution. If the author does not have the financial resources to cover these charges, a waiver of the charges may be requested of the editor. Articles are selected without regard to authors' ability to pay page charges. In cases in which an institution is responsible for the charges, authors are urged to submit an official purchase order to cover the charges. Please list the vendor as The University of Chicago Press.

Editorial Board

Editor in Chief

Walter M. Fitch
Department of Ecology and Evolutionary Biology
University of California, Irvine
716 Engineering Building
Irvine, California 92717

Managing Editor

Masatoshi Nei
Institute of Molecular Evolutionary Genetics
328 Mueller Lab
Pennsylvania State University
University Park, Pennsylvania 16802

Associate Editors

Wesley M. Brown
Department of Biological Sciences
University of Michigan
Ann Arbor, Michigan 48109-1048

Michael Bulmer
Department of Biological Sciences
Rutgers University
Piscataway, New Jersey 08855-1059

Brian Charlesworth
University of Chicago
1103 East 57th Street
Chicago, Illinois 60637

Richard B. Flavell
Director, John Innes Institute
John Innes Centre for Plant Science Research
Colney Lane
NORWICH
NR4 7UH
England

Barry G. Hall
Department of Biology
University of Rochester
Hutchison Hall
Rochester, New York 14726

Jan Klein
Max Planck Institut für Biologie
Abteilung Immungenetik für Biologie
Corrensstrasse 42, Tübingen 1
Federal Republic of Germany D-7400

Richard K. Koehn
Department of Ecology and Evolution
State University of New York at Stony Brook
Stony Brook, New York 11794

Martin Kreitman
Department of Ecology and Evolutionary Biology
Princeton University
Princeton, New Jersey 08544-1003

Alan Weiner
Department of Molecular Biophysics and Biochemistry
Yale University
333 Cedar Street, SHM-C16
New Haven, Connecticut 06510

Elizabeth A. Zimmer
Laboratory of Molecular Systematics
Museum Support Center
Smithsonian Institution
Washington, D.C. 20560
A Selection of Recent Articles

*In vivo* destruction of canine lymphoma mediated by munne monoclonal antibodies. Z. STEPLEWSKI, C. ROSALES, K.A. JEGLUM, J. MCDONALD-SMITH (Philadelphia, PA, USA)


Nude mouse xenografts as *in vivo* models for lung carcinomas. G. BEPLER, K. NEUMANN (Marburg, FRG)

MIIC class I antigen expression and alterations induced by rIFN-γ in tumor hybrid cell immunogenicity. A.F. GOGUEL, F. LE BORGNE, B. BAUDE, M.L. CANAVATE, S. TLOUZEAU, M.L. DERHY, G. LESPINATS (Villejuif, France; Vizcaya, Spain)

Radiosensitivity of different human tumor lines grown as xenografts determined from growth delay and survival data. J.H.M. SCHWACHÖFER, J. HOGENHOUT, H.B. KAL, J. KOEDAM, H.P.N. VAN WEZEL (Nijmegen; Rijswijk, The Netherlands)

Effect of total parenteral nutrition on histamine in the rat brain and other tissues. B.J. HINE, N.L. KEIM, L.R. HEGSTRAND (Madison, WI, USA)

Evaluation of N-4-(hydroxycarbophenyl) retinamide as a cancer prevention agent and as a cancer chemotherapeutic agent. J. HAN, L. JIAO, Y. LU, Z. SUN, Q.M. GU, K.J. SCANLON (Beijing, China; Duarte, CA, USA)

Modifying factors in prenatal carcinogenesis. V. ALEXANDROV, C. AIELLO, L. ROSSI (Leningrad, USSR; Genoa, Italy)

Organization and modulation of a novel nuclear phosphoprotein during rat spermatogenesis. E.C. CHEW, A.S.O. YEUNG (Shatin, Hong Kong)

Murine mycotic placentalitis produced by intravenous inoculation of conidia from *Aspergillus fumigatus*. H.E. JENSEN, J. HAU (Frederiksberg; Copenhagen, Denmark)

Critical period of induction by tamoxifen of genital organ abnormalities in female mice. S. IRISAWA, T. IGUCHI (Yokohama, Japan)

Immunological reactivity to a mycobacterial fraction is associated with nonspecific suppression of immunological responsiveness in vivo. D. HALPERIN, C. REUBEN, G. RASHID, S. BEN-EFRAIM, D.W. WEISS (Tel-Aviv; Jerusalem, Israel)

Segmental liver resection with linear stapling device. An experimental study on pigs. T. ZILLING, B.S. WALther, T. HOLMEN (Lund, Sweden)

Acute and chronic effects of diethylstilbestrol on serum levels of alpha melanocyte stimulating hormone and prolactin in the hamster. R.N. KURL, P.H. NAYLOR, J.M. LORING, R.B. TODD, C.A. VILLEEE (Boston, MA, USA)

*In vivo* effects of recombinant human and murine interleukin-1α on murine hematopoiesis. C.S. JOHNSON, P.G. BRAUNSCHWEIGER, P. FURMANSKI (Denver, CO, USA)

Synergistic antitumor effect of interleukin-2 and irradiation on pancreatic cancer in Syrian golden hamsters. T. MORIAI, I. MAKINO, Y. KIKUCHI (Asahikawa, Japan)

**General Policy:** *IN VIVO* is a multi-disciplinary journal designed to bring together original high quality works and reviews on biomedical research involving experimental animals. Articles are drawn from scientific areas including oncology, pharmacology, chemotherapy, immunology, radiology, toxicology, genetics, cytology, endocrinology, nutrition. The principal aim of *IN VIVO* is to provide for the prompt publication of accepted articles, generally within 1-2 months from final acceptance.

**Publication Data:** *IN VIVO* is published bimonthly. Each annual volume contains six issues. Annual Authors and Subject Indexes are included in the sixth issue of each volume. Manuscripts should be submitted to: Dr. J.G. Dezinassis, Managing Editor, *IN VIVO*, Editorial Office, 5 Argyropoulou Street, Kato Patissia, Athens GR 111 45, Greece.

Articles in *IN VIVO* are regularly indexed in bibliographic services, including Biological Abstracts, Chemical Abstracts, BIOSIS, Excerpta Medica, CABS Database, CANCERLIT Database, SUBIS, Reference Update-RIS.

**Special Offer:** The first author of each accepted article will be entitled to a free one-year subscription to the journal.


**Editorial Office:** 5 Argyropoulou Street, Kato Patissia, Athens, GR-111 45, Greece. Telephone and Fax 2016380.
Blackwell Scientific Publications are proud to announce the launch of a prestigious, international, primary journal aimed at the interface between molecular biology and the ecological sciences.

*Molecular Ecology* will publish the highest quality papers in the newly-emerging and rapidly developing field of molecular ecology, with particular emphasis on natural populations and environments, including the release of genetically engineered organisms. Submitted papers will be considered by a distinguished international panel of editorial advisors, and editorial and production procedures will ensure rapid publication. The remit of *Molecular Ecology* will be to provide a vehicle for the publication of the best fundamental research findings within the following, and related fields:

- Molecular Population Biology
  - population and evolutionary genetics
  - behavioural ecology
  - conservation biology
- Molecular Environmental Genetics
  - population ecology and gene flow
  - ecological aspects of environmental release including recombinant organisms
  - genetic exchange within the natural environment
  - competitiveness and species invasion
- Molecular Adaptation
  - genetic differentiation and physiological adaptation
  - environmental effects on gene expression
- Technological Developments in Molecular Ecology
  - molecular techniques for species identification
  - development of novel probes
  - sequences and primers for population studies

**Receiving Editors**

Dr Terry Burke  
Department of Zoology, University of Leicester,  
Leicester LE1 7RH, UK

Dr Ray Seidler  
2450 SW 35th Street, Corvalis, Oregon  
97333, USA

**Co-ordinating Editor**

Professor Harry Smith  
Department of Botany, University of Leicester,  
Leicester LE1 7RH, UK

*Molecular Ecology* will be published quarterly from 1992. Subscription prices will be £30.00 for institutions and £25.00 for individuals.

[ ] Please reserve my free specimen copy of Issue One

[ ] Please send me full author instructions

Name  
Address  
Post/Zipcode

Please return this form to Anna Rivers at the address below.
PERSPECTIVES IN BIOLOGY AND MEDICINE

Reasoned essays, critical assessments, and autobiographical accounts—engaging scientists, physicians, and students in and outside their specialties.

Perspectives in Biology and Medicine was founded in 1957 to stimulate original thought in the life and medical sciences. Today, Perspectives continues to present lively, probing articles written for readers representing an array of specializations.

Each issue treats a variety of topics with an emphasis on interpretive essays that convey the author's opinion—his or her perspective on a subject of interest to physicians and scientists. You can expect to find, for example, personalized views of biomedical history, speculative pieces on the evolutionary biology of schizophrenia, and humorous pieces on the dilemmas of retaining the tenure system for clinical faculties.

You'll also find articles that focus on the relationship of medicine to society. Recent issues have included discussions of:

- euthanasia
- abortion
- do-not-resuscitate
- fee-for-service research
- policy
- medicine and public interest

Perspectives is not the kind of journal you read in one day; while relaxing, you take in an article here and ponder another there.

Edited by Richard L. Landau
Sponsored by The Division of Biological Sciences, The University of Chicago
Published quarterly by The University of Chicago Press
Published monthly for the American Society of Human Genetics by the University of Chicago Press

Since 1949 The American Journal of Human Genetics has provided an authoritative record of research and review relating to human heredity. The Journal focuses on applications of genetic principles in medicine, psychology, physical anthropology, and social services and on relevant areas of molecular and cell biology. Your annual subscription to AJHG, including approx. 2,300 pages, features

- original research articles, all peer-reviewed, most appearing 11-13 weeks after final acceptance
- critical commentaries, review articles, and book reviews
- a quarterly section on Human Genetics Education, including reports on instructional innovations and reviews of teaching materials
- announcements of conferences and employment/fellowship opportunities

Researchers and clinicians in the biomedical sciences will find in AJHG the rapid progress and scientific rigor that characterize this field.

Editor: Charles J. Epstein, M.D., University of California, San Francisco

One-year subscription rates: Nonmember individuals and institutions $250.00. Student subscriptions available only as members of the American Society of Human Genetics (See below). Outside USA add $24 for surface postage. Subscriptions entered on a calendar-year basis only. Visa and MasterCard payments accepted. To order send check, purchase order, or complete credit card information to The University of Chicago Press, Journals Division, Dept. SF1FA, P.O. Box 37005, Chicago, IL 60637.

American Society of Human Genetics: Membership in the society includes a subscription to AJHG. For membership information, please write to the American Society of Human Genetics, Administrative Office, 9550 Rockville Pike, Bethesda, MD 20814.
III CONGRESO LATINOAMERICANO DE HERPETOLOGIA
(Third Latinamerican Congress of Herpetology)

To Be Held In Campinas, São Paulo, Brasil
December 12-18, 1993

For Further Information Write:

Comissão Organizadora III CLAH
Departamento de Zoologia, IB
Universidade Estadual de Campinas
13081 - Campinas, São Paulo
Brasil
Journals from
The University of Chicago Press

Send for our FREE Journals Catalog today

At the forefront of academic publishing since 1891, The University of Chicago Press brings you the highest standards of scholarship.

SOCIAL SCIENCES
Adolescent Psychiatry
American Journal of Sociology
Child Development
Child Development Abstracts and Bibliography
Current Anthropology
Economic Development and Cultural Change
International Annals of Adolescent Psychiatry
Isis
Journal of British Studies
Journal of Consumer Research
Journal of the History of Sexuality
Journal of Labor Economics
The Journal of Law & Economics
The Journal of Legal Studies
The Journal of Modern History
Journal of Political Economy
Law & Social Inquiry
Monographs of the Society for Research in Child Development
Ocean Yearbook
Osiris
Public Opinion Quarterly
Signs: Journal of Women in Culture and Society
Social Service Review
The Supreme Court Review
Technology and Culture

HUMANITIES
Classical Philology
Crime and Justice
Critical Inquiry
Ethics
History of Religions
International Journal of American Linguistics
Journal of Near Eastern Studies
The Journal of Religion
The Library Quarterly
Modern Philology
Winterthur Portfolio: A Journal of American Material Culture

BIOLOGICAL AND MEDICAL SCIENCES
The American Journal of Human Genetics
The American Naturalist
International Journal of Plant Sciences
Clinical Infectious Diseases
The Journal of Infectious Diseases
Molecular Biology and Evolution Perspectives in Biology and Medicine
Physiological Zoology
The Quarterly Review of Biology

EDUCATION
American Journal of Education
The Bulletin of the Center for Children's Books
Comparative Education Review
The Elementary School Journal

For a complete catalog or brochures on individual journals—write to: Sandra Willis, Circulation Department, The University of Chicago Press, 5720 S. Woodlawn Avenue, Chicago, IL 60637 USA
Forthcoming

Positive Selection and Interallelic Recombination at the Merozoite Surface Antigen–1 (MSA-1) Locus of Plasmodium falciparum

Austin L. Hughes

Population Growth Makes Waves in the Distribution of Pairwise Genetic Differences

Alan R. Rogers and Henry Harpending

The Involucrin Gene of Old-World Monkeys and Other Higher Primates: Synapomorphies and Parallelisms Resulting from the Same Gene-altering Mechanism

Philippe Djian and Howard Green

Mitochondrial DNA Evolution at a Turtle’s Pace: Evidence for Low Genetic Variability and Reduced Microevolutionary Rate in the Testudines

John C. Avise, Brian W. Bowen, Trip Lamb, Anne B. Meylan, and Eldredge Bermingham

A Stochastic Analysis of Three Viral Sequences

G. A. Watterson

Phylogeny of the Bovidae (Artiodactyla, Mammalia), Based on Mitochondrial Ribosomal DNA Sequences

John Gatesy, Deborah Yelon, Rob DeSalle, and Elisabeth S. Vrba

Major-Histocompatibility-Complex DRB Genes of a New-World Monkey, the Cotton-top Tamarin (Saguinus oedipus)

Blazenka Grahovac, Werner E. Mayer, Vladimir Vincek, Felipe Figueroa, Colm O'hUigin, Herbert Tichy, and Jan Klein

Selection of Laboratory Wild-Type Phenotype from Natural Isolates of Escherichia coli in Chemostats

Riitta Mikkola and C. G. Kurland

A Comprehensive Study of Genic Variation in Natural Populations of Drosophila melanogaster. VII. Varying Rates of Genic Divergence as Revealed by Two-dimensional Electrophoresis

Shanta Thomas and Rama S. Singh

The Cytochrome b and ATPase Genes of Honeybee Mitochondrial DNA

R. H. Crozier and Y. C. Crozier

Geographic Mitochondrial DNA Variation in the Rock Hyrax, Procavia capensis

Paulette Prinsloo and Terence J. Robinson

The Consistency of Several Phylogeny-Inference Methods under Varying Evolutionary Rates

Ronald W. DeBry

Comparative Structural Analysis of the Transcriptionally Active Proopiomelanocortin Genes A and B of Xenopus laevis

Peter M. T. Deen, Marion J. M. Bussemakers, Dick Terwel, Eric W. Roubos, and Gerard J. M. Martens

An Association between ADH Protein Levels and Polymorphic Nucleotide Variation in the Adh Gene of Drosophila melanogaster

Paul Matthew, Alex Agrotis, Andrea C. Taylor, and Stephen W. McKechnie


Austin L. Hughes
Molecular Biology and Evolution
Volume 9, Number 2, March 1992

Original Articles

193  Sequences of Primate Insulin Genes Support the Hypothesis of a Slower Rate of Molecular Evolution in Humans and Apes than in Monkeys
Susumu Seino, Graeme I. Bell, and Wen-Hsiung Li

204  Mitochondrial DNA–like Sequence in the Nuclear Genome of an Akodontine Rodent
Margaret F. Smith, W. Kelley Thomas, and James L. Patton

216  Coevolution of the Vertebrate Integrin α- and β-Chain Genes
Austin L. Hughes

235  Phylogenetic Position of Phylum Nemertini, Inferred from 18S rRNA Sequences: Molecular Data as a Test of Morphological Character Homology
J. McClintock Turbeville, Katharine G. Field, and Rudolf A. Raff

250  Molecular Relationships between Alcohol Dehydrogenase Null-Activity Alleles from Natural Populations of Drosophila melanogaster
John B. Gibson, Ann Verona Wilks, and Alex Agrotis

261  The Drosophila subobscura Adh Genomic Region Contains Valuable Evolutionary Markers
Gemma Marfany and Roser González-Duarte

278  Two Ascomycete Classes Based on Fruiting-Body Characters and Ribosomal DNA Sequence
Mary L. Berbee and John W. Taylor

285  Molecular Evolution of the Fungi: Relationship of the Basidiomycetes, Ascomycetes, and Chytridiomycetes
Barbara H. Bowman, John W. Taylor, Alan G. Brownlee, Jean Lee, Shi-Da Lu, and Thomas J. White

297  The Two Similarly Expressed Genes Encoding U3 snRNA in Schizosaccharomyces pombe Lack Introns
David A. Selinger, Gregory L. Porter, Patrick J. Brennwald, and Jo Ann Wise

309  Character Analysis and the Integration of Molecular and Morphological Data in an Understanding of Sand Dollar Phylogeny
Charles R. Marshall
Rapid Surveying of DNA Sequence Variation in Natural Populations
Enrique P. Lessa

Sequence Instability and Functional Inactivation of Murine Y Chromosomes Can Occur on a Specific Genetic Background
Ferez S. Nallaseth

Letters to the Editor

The Number of Replications Needed for Accurate Estimation of the Bootstrap P Value in Phylogenetic Studies
S. Blair Hedges

Substitution Bias, Weighted Parsimony, and Amniote Phylogeny as Inferred from 18S rRNA Sequences
Charles R. Marshall

18S rRNA Sequences and Amniote Phylogeny: Reply to Marshall
S. Blair Hedges and Linda R. Maxson

Commentary on Letters of Marshall and of Hedges and Maxson
Walter M. Fitch