Selection on Silent Sites in the Rodent H3 Histone Gene Family

Ronald W. DeBry* and William F. Marzluff†

*Department of Biological Science, Florida State University, Tallahassee, Florida 32306-2043, and †Program in Molecular Biology and Biotechnology, University of North Carolina, Chapel Hill, North Carolina 27599

Manuscript received November 22, 1993
Accepted for publication May 6, 1994

ABSTRACT

Selection promoting differential use of synonymous codons has been shown for several unicellular organisms and for Drosophila, but not for mammals. Selection coefficients operating on synonymous codons are likely to be extremely small, so that a very large effective population size is required for selection to overcome the effects of drift. In mammals, codon-usage bias is believed to be determined exclusively by mutation pressure, with differences between genes due to large-scale variation in base composition around the genome. The replication-dependent histone genes are expressed at extremely high levels during periods of DNA synthesis, and thus are among the most likely mammalian genes to be affected by selection on synonymous codon usage. We suggest that the extremely biased pattern of codon usage in the H3 genes is determined in part by selection. Silent site G + C content is much higher than expected based on flanking sequence G + C content, compared to other rodent genes with similar silent site base composition but lower levels of expression. Dinucleotide-mediated mutation bias does affect codon usage, but the affect is limited to the choice between G and C in some fourfold degenerate codons. Gene conversion between the two clusters of histone genes has not been an important force in the evolution of the H3 genes, but gene conversion appears to have had some effect within the cluster on chromosome *13*.

THE replication-dependent H3 histone genes of the rodent genus *Mus* form a highly homogeneous, medium-sized multigene family consisting of approximately 15–20 copies on two chromosomes (MARZLUFF and GRAVES 1984; GRAVES *et al.* 1985). The H3 protein is extremely conservative at the protein sequence level, which accounts for much of the nucleotide sequence similarity observed between gene copies. However, even at silent positions four H3 genes from *Mus musculus* averaged 91.3% identical (TAYLOR *et al.* 1986). Neutral sequences are often very similar among copies in multigene families due to frequent gene conversion and unequal exchange, mechanisms that act to homogenize a gene family by repeatedly duplicating and deleting large blocks of sequence (HOOD *et al.* 1975; SMITH 1976; DOVER 1982). These mechanisms act without regard for coding capacity or functional importance, and thus can maintain sequence similarity among copies in the absence of any selective force. For example, the nontranscribed spacer of ribosomal RNA genes are highly homogeneous in many species (HILLIS and DIXON 1991).

However, the Mus replication-dependent H3 genes show an extreme pattern of codon-usage bias (TAYLOR *et al.* 1986), suggesting the possibility that some or all of the silent sites also may be influenced by selection. Selective differences between synonymous codons could be due to differences in either the efficiency or accuracy of the translational process (BULMER 1991), or to requirements of mRNA secondary structure (HUYNEN *et al.* 1992). In either case, selection coefficients against a particular suboptimal codon are likely to be very small, so extremely large effective population sizes would be needed for selection to overcome the stochastic effects of drift. For this reason, codon selection was originally thought to be limited to unicellular organisms such as bacteria and yeast (GOUY and GAUTIER 1982; IKEMURA 1985; SHARP and LI 1986). Recently, however, selection has also been invoked to explain patterns of codon bias seen in Drosophila, despite the much smaller effective population sizes compared to unicellular organisms (SHIELDS *et al.* 1988; KLIMAN and HEY 1993; MORIYAMA and HARTL 1993). Effective population sizes in rodents are likely to be one or two orders of magnitude smaller than those of most Drosophila species (NEI and GRAUR 1984), increasing the likelihood that drift and mutation pressure will be the primary determinants of codon usage. Codon usage in mammalian genes usually is thought to be determined primarily by differential mutation pressure, because silent site base composition is generally correlated with the base composition of surrounding noncoding sequences (BERNARDI *et al.* 1985; AOTA and IKEMURA 1986; FILIPSKI 1987). An additional consideration for the replication-dependent histone genes is that evolutionary mechanisms other than selection can operate in multigene families (ARNHEIM 1983). This means that selection may be less effective on a particular site in a multigene family than on an equivalent site in a single-copy gene (HOOD *et al.* 1975; OHTA 1980).

Here, we examine sequences of replication-dependent H3 genes from three species, including two murid rodents and one cricetid rodent. We examine patterns of codon usage and the relationship between silent site base composition and flanking sequence base composition to determine if mutation pressure alone can account for the observed codon-usage bias. We also infer historical relationships among the coding sequences within and between species, to assess whether gene conversion alone can account for the high level of homogeneity in the H3 gene family.

Structure of the rodent H3 histone gene family: The core histones (H2A, H2B, H3 and H4), along with the H1 linker histone, help form the basic structural unit of the nucleosome. There are several variant forms of most histone proteins in many organisms, which usually are expressed only at some stages of development or in certain tissue types. Here, we consider only the replication-dependent histone genes (ZWEIDLER 1984). These are the major forms that are expressed at extremely high levels only during S phase of the cell cycle throughout most of the life of the organism. At the amino acid sequence level, these are among the most conservative of all proteins (DELANGE *et al.*, 1969; PATTHY *et al.* 1973). The mammalian replication-dependent H3 genes are small [411 nucleotides (nt)] and compact, with no introns and short 5′ (<40 nt)-and 3′ (<60 nt)-untranslated flanking sequences. They are not polyadenylated; instead the mature mRNA ends 3′ in a conserved stem-loop structure (HENTSCHEL and BIRNSTIEL 1981) formed by a posttranscriptional processing reaction involving U7 snRNA, which forms a duplex with a conserved site several bases downstream from the stem-loop (MOWRY and STEITZ 1987).

Unlike the histone genes of many invertebrates and some vertebrates [*e.g.*, Notophthalamus (STEPHENSON *et al.* 1971) and Xenopus (ZERNICK *et al.* 1980)], the replication-dependent histone genes in Mus are not organized as tandem repeats. Instead, 15–20 genes for each of the five proteins are jumbled together in an apparently random order and orientation, with apparently random intervening sequences that range from only a few hundred base pairs to well over 15 kb in length (SEILER-TUYNS and BIRNSTIEL 1981; SITTMAN *et al.* 1983; GRUBER *et al.* 1990; S.-F. WANG, W. F. MARZLUFF and R. W. DEBRY, unpublished data). A similar organization is found in the few birds that have been examined (ENGEL and DODGSON 1981), while sea urchins have both a large, tandemly repeated array that is expressed early in development and a smaller, jumbled cluster that is expressed later (MAXSON *et al.* 1983). In addition to the large cluster, which has been mapped to *M. musculus* chromosome *13*, there is another, smaller cluster of replication-dependent histone genes on *M. musculus* chromosome *3* (GRAVES *et al.* 1985). So far, only one H3 and one H2A gene have been found in this cluster, and it appears that no additional copies of the H3 gene are found on chromosome *3*. Interestingly, this single H3 gene produces approximately 40% of the total H3 mRNA found in S-phase cells, while each of the chromosome *13* genes makes approximately 5% of the total H3 mRNA (GRAVES *et al.* 1985).

Other than two short segments of conserved 3′-flanking sequence (the stem-loop and U7-binding region) and a few similarly small 5′-promoter sequences (a TATAA box and one or more CCAAT boxes), the flanking sequences of each *M. musculus* H3 gene so far examined are unique to that particular gene, beginning immediately 5′ of the start codon and immediately 3′ of the stop codon (TAYLOR *et al.* 1986; this study). One conclusion that can be drawn from this observation is that unequal exchange does not play a significant role in maintaining sequence homogeneity among the H3 coding sequences. Unequal exchange acts on all sequences between the exchange sites, whether coding or noncoding, as seen in the non-transcribed spacer region of tandemly repeated ribosomal RNA genes. This feature of rodent histone genes allows examination of the effects of gene conversion in a setting where the conversion events are not confounded by duplications and deletions produced by unequal exchange.

Gene conversion does occur between rodent replication-dependent histone genes, and therefore is a potential mechanism of homogenization. LIU *et al.* (1987) sequenced two copies of the H2A gene, one of which is clearly a pseudogene, as it lacks the first 9 and the last 3 amino acids, and the promoter region. Nonetheless, a 350-bp stretch including virtually all of the rest of the coding region shows only a single base change from the adjacent functional H2A gene. It is possible that one gene might be the target of multiple conversions, which could make it difficult to detect a particular conversion event. However, *in vivo* experiments using plasmid constructs showed that most conversion events involved at least 400 bp of contiguous sequence (LISKAY *et al.* 1987), which is nearly the same size as the entire H3 coding region. Therefore, with the exception of the ends of the coding sequence, it is likely that only the most recent conversion event would be observed.

MATERIALS AND METHODS

The *Nco*I-*Pst*I fragment (282 nt) of the Mm614 H3 gene was radioactively labeled using random hexanucleotide primers (FEINBERG and VOGELSTEIN 1983) and used to screen plaques of a *Mus pahari* EMBL3 genomic library. Positive clones were rescreened until pure cultures were obtained. H3 genes were subcloned into pGEM vectors and sequenced on an ABI 373A automatic sequencer. The hamster gene (ARTISHEVSKY *et al.* 1987) was subcloned into pGem vectors for sequencing. The *M. musculus* genes, from laboratory strain BALB/c, were originally reported by TAYLOR *et al.* (1986) and GRUBER *et al.* (1990). GenBank accession numbers for sequences used in this study are: X80324–X80327, X80330, X16148, M32462, M32459, M32460. Flanking sequences were aligned using the ESEE computer program (written by E. CABOT). Phylogenetic analyses of the codon regions were performed using PAUP 3.1 (SWOFFORD 1993).

SHIELDS *et al.* (1988) proposed the "Scaled" χ^2 (a χ^2 calculated on the deviation from equal usage of all synonymous codons, divided by the total number of codons excluding Trp and Met) as a measure of codon-usage bias that is independent of gene length, and one that can be used to compare codon usage across genes and across taxa. For the two amino acids encoded by six synonymous codons (Leu and Arg), each codon is expected to be used at $\frac{1}{6}$ of the positions. One drawback of this formulation is that it is difficult to determine expected codon usage if the equilibrium base frequency is not 0.25 for each base.

RESULTS

Codon-usage bias

Replication-dependent H3 gene sequences: Sequences for four H3 genes from *M. pahari*, as well as one H3 gene from hamster were aligned to five previously published H3 sequences from *M. musculus* (TAYLOR *et al.* 1986; GRUBER *et al.* 1990; Figure 1). Two of the *M. pahari* genes are of the H3.1 subtype (cysteine at residue 96) and two are of the H3.2 subtype (serine at residue 96); the hamster gene is an H3.2 subtype. These are the only two known protein variants for the replication-dependent H3 genes in mammals. Both protein subtypes have been found in all mammals examined, although their relative abundance varies from about 80% H3.2 in rodents to about 80% H3.1 in primates (MARZLUFF 1986). All of the genes included in this study code for functional protein; there are no insertions, deletions, frameshifts or amino acid substitutions other than at residue 96. The flanking sequences of all these genes contain previously described conserved promoter sequences in the 5′ region and all the predicted mRNAs have the conserved 3′ stem-loop structure and U7 snRNA-binding sequence that are required for proper 3′ end formation.

The *M. pahari* genes follow the pattern noted by TAYLOR *et al.* (1986) for *M. musculus*, where each gene has unique 5′- and 3′-flanking sequences except for the relatively small functional elements (Figure 1). This extreme divergence in flanking sequence between genes within a species can be used to identify orthologous genes between species. Examination of flanking sequences reveals that the 5′ sequences from the hamster gene and the *M. pahari* gene Mp2.3 are both much more similar to the 5′ sequences from the *M. musculus* Mm614 gene than to any of the other 5′-flanking sequences, indicating that these three genes are orthologous (Figure 2A). The similarity between the hamster and the two Mus sequences continues in the 5′ direction through the H2A gene located approximately 1 kb from the H3 gene (R. W. DEBRY and W. F. MARZLUFF, unpublished results). The 3′ sequences from the two species of Mus are also quite similar to each other, but the hamster 3′ sequence is much more divergent (Figure 2B). Still, the hamster 3′ sequence shares a CTTCCCGG sequence immediately upstream from the U7-binding site that is not found in any flanking sequences besides Mm614 and Mp2.3. It may be that the 3′ sequences diverge faster than the 5′ sequences, possibly due the presence of unrecognized 5′ regulatory elements, or it is possible that a large sequence rearrangement has occurred 3′ from the U7-binding sequence since the divergence between Mus and hamster.

Mm614 is the single, highly expressed H3 gene found on *M. musculus* chromosome 3 (GRAVES *et al.* 1985). In *M. musculus*, all the known replication-dependent H3 genes except for Mm614 are located on chromosome *13* (GRAVES *et al.* 1985), and we will assume that all of the *M. pahari* genes reported here except Mp2.3 are located on the *M. pahari* equivalent of *M. musculus* chromosome *13*. Among the chromosome *13* genes, we have found one additional pair of orthologous genes between the two species of Mus: both the 5′- and 3′-flanking sequences of Mm291 and Mp1.5 are highly similar (Figure 1). As further evidence that these two genes are orthologous, the plasmid insert containing the 5′ half of Mp1.5 also contains an H2A and an H2B gene that have flanking sequences very similar to those of genes found in the same relative position in *M. musculus* (R. W. DEBRY and W. F. MARZLUFF, unpublished results).

Pairwise similarities at the 163 silent codon positions (133 silent third positions and 30 silent first positions at alanine and leucine codons) average 90.2% (range 84.0–97.5%) across all the *M. musculus* and *M. pahari* genes (Table 1). Within *M. musculus*, pairwise similarities average 91.3%, while the *M. pahari* genes are identical at an average of 89.5% of the silent sites.

The replication-dependent H3 genes show an extreme pattern of codon-usage bias. Among the chromosome *3* genes, only 1 out of a total of 399 silent sites are occupied by an A. Across all silent sites, the chromosome *3* genes are over 96% G/C, while the eight chromosome *13* genes from *M. musculus* and *M. pahari* average approximately 91% G/C.

Scaled χ^2 values: We used the scaled χ^2 (SHIELDS *et al.* 1988) to provide a metric of codon-usage bias that can be compared across genes and taxa. Scaled χ^2 values range from 1.55 to 1.63 for the chromosome *3* genes, and from 1.31 to 1.69 for the chromosome *13* genes (Table 2). All of these values are higher than those for any of the Drosophila genes examined by SHIELDS *et al.* (1988) or MORIYAMA and HARTL (1993).

Codon usage compared to flanking sequence composition: In mammals, high G + C content at either silent sites or third position sites (and thus high codon-usage bias) is correlated with high G + C content in flanking sequences and introns (BULMER 1987a; WOLFE *et al.* 1989). High G + C content genes are usually embedded within high G + C isochores. This pattern, coupled with the relatively small effective population size of most mammals, provides evidence that selective differences between synonymous codons do not produce the codon-usage patterns observed in most mammalian genes.

If a mutation-bias mechanism is all that is needed to explain the codon usage bias in the replication-dependent histone genes, then the extremely high G + C content of the silent positions predicts that these

```
Mm614                         GGGACCCGCCCTCTACGGGCAGCTGCCAGACGTGGCTCCACCCTCGCAGAACGTTGGCAAGT -234
Mp2.3                         AGCCGGACCCGCCCTCTACGGGCAGCTGCCAGACGTGGCTCCACCCTCGCAGAAGGTTGGCG -234
Mm221-1                       TTTAAGAAAGCTCGGGTGTGCCAGACTAAAACATGAGTGTAGCAGTCTGTACTAGCAGGAGA -234
Mm221-2                       TGGTATCGCCAAAATCTACATAGTATCTCTTATTAAAATGTTTTGACGAAAATGTAACAATT -234
Mm291                         AGTCACTACCTACCTTTTACACATCTTTTCATTTAATCTGAAGTAGGAAAAAGAAAGTAAGA -234
Mp1.10                        TTCAATGGCACTAAGTCCAAGCCGCAATGTGTTTGGAAGCAGAAGTGCTGTAGAAGATGACA -234

Mm614         GACCGGGCGCGGGGCTGGACGTGGGGGGCGGGTGGGGGCGGGTCAGGTGGCGTGGCGGGCCCGAGCCAATGGGCGAGGT -156
Mp2.3         ATCGGCCGGGCGCCAGGCTGGACCTGGGGGGCGGGGGGCGGGTCTGGAGGCGTGACGGGCCCGAGCCAATGGGCGAGGT -156
Mm221-1       CGTGTGCAGGAGTTAACCAATCGGGTGTGCAGGAGTTAACCAATCACCACTTGAATTCTCAGCCAATAGGACTACTGCG -156
Mm221-2       AGGTACTTAATGGTAACAATGTGGCTGAGGAAGCAATAGTTAACAAAGAGGAGCTAAGCTATGCAACAAACCAGATTTC -156
Mm291         AAGTGCTTTTAATAAAAATTACATTGATTTAAACGAACCTGCGGCGGGATTTTCTTTTCAATGAGGAAATTGCACACTT -156
Mp1.10        GCACTGGAATCTTATTCCATTTTCCTTAAAACACAGACATAGAAAAATAAGTGCGGCTGCAGACTACCTTAGCCCCAGA -156

Mm614         CGGGGCCGGTGACGCCACGGCCAATGGCGCGGCAGCGCGGGAGTTTCAAGTCGCTGTCTCCGCCCCGCCGCGGGGAAGA  -78
Mp2.3         CGGTGCCTGTGACGTCACGGCCAATGGCGCGGCAGCGCGGGAGTTTCAAGTCGCTGTCTCCGCCCCGCCGCGGGGAAGA  -78
Hamster                                                                       GGTGGGGGGGGAAGA  -78
Mm53                  CAGGATTTAGAAGCAGAGGCTGACCAATCCCAACAAAGCGCGGCCCTTTGAATGTTCTTCGGTCCAATAG  -78
Mm221-1       CGGGACACTTGAAAAGCAGACACGCCTATCAGGATGCTTTCTCGGTGGGAAGGAGGGGTACGAGCGCGGGTACGTGTGT  -78
Mm221-2       TATTGGTCACAAATTTGAAGTTGAGACCTGTTATCCAATTACCAAGTACTTCCGCATACATCATCATAGGCATTTGAAG  -78
Mm291         GGTTAAACCAAGTTCAGACTGCGAAAACAAAAGGACTCACCAGCCAATTAAGTTGATCTGGCAGCCATTTTGACCCAAT  -78
Mp1.5                                                                           ATTTTGACCCAAT  -78
Mp1.10        TAACCAGCAACTGTAGTCCTAAATAAACCAATCAGAGTCTTAAACGTCACAGATAACCAGTATTTTCATCCAATCACTA  -78

Mm614         CTGCGCCTATAAAGGCGGCCGGCTCGGGCCGGTATCAGGTCCCCGAGTGCTCCTCGTTTGGGCGTCTTCCGTCTTCGCC   -1
Mp2.3         CTGAGCCTATAAAGGCGGCCGGCTCGGGCCGGTGTCAGGTCCCCGTGTGCTCCTCGTCTGGGTGTCTTCCGTCTCCGCC   -1
Hamster       CTGTTCCTATAAAGACGGTCGNCTGCGACCTGGGTCTAGTCCCCGTGTGCTCCTCGTCTTGGTGACTTCCGTCTCCGCC   -1
Mm53          CGGATAGTCTGATTGTATAAAAGGTGGACAGCGCCTTGCAGCTCACTATAGTGTCAGTCTATTTTCCCTTGTTAAAGTC   -1
Mm221-1       TGCGCGTGTGCGACGCAAGCGTACTTAAAGGCCAAAGTGCGCTACTTAGGTATCTCACTTTTCCCTACGGTTACTTGCC   -1
Mm221-2       ATTTCAACCAATCAGGAGCATGTTCCTTCTATAAAGGAACCCAGAACCTAACCTCTGCATTTCCTATTTCTTTGTAGAA   -1
Mm291         CAGAACTCGGCCGTCTGTATAAATTTTGGTGGTTGAAGCTTTCCCTCCATCACTTTGCTTTGGAAGCTCGGGTGTTACC   -1
Mp1.2                                          TTTGCAGCGCACTGTAGTGTTAGTTGTTTTTCAGTCTTTACAGTA   -1
Mp1.5         CAGAACTCGCCCATCTGTATAAAATTTTGGTGGTGGANCTTTCCCGCTATCACTTCGCTTTGGAAGCTAAAGTGTTGCT   -1
Mp1.10        TTCTTGGACACTATAAATAGTAGTTCTGAGCTTCTCACTTCCATGTTCCTAGTCCTCAGCCGCCTTTCAGGTCCTTGCA   -1

Mm614         ATG GCC CGT ACG AAG CAG ACC GCC CGC AAG TCC ACC GGC GGC AAG GCC CCG CGC AAG CAG   60
Mp2.3         ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ..G ... ... ... ...   60
Hamster       ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...   60
Mm53          ... ..T ... ..T ... ... ... ..T ..T ... ... ... ..T ... ... ... ... ..T ... ...   60
Mm221-1       ... ..T ... ..T ... ... ... ..T ... ... ..T ... ... ... ... ... ... ... ... ...   60
Mm221-2       ... ..T ... ..T ... ... ... ..T ... ... ... ..T ... ... ... ... ... ... ... ...   60
Mm291         ... ..T ... ..T ... ... ... ..T ... ... ... ... ... ... ... ... ... ... ... ...   60
Mp1.2         ... ..T ... ..A ... ... ... ..T ... ... ... ... ... ... ... ..T ... ... ... ...   60
Mp1.5         ... ..T ... ..T ... ... ... ..T ... ... ... ... ... ... ..A ..T ... ... ... ...   60
Mp1.10        ... ..T ..C ..A ... ... ... ..T ... ... ... ... ... ... ... ... ... ... ... ...   60

Mm614         CTG GCC ACC AAG GCC GCC CGC AAG AGC GCC CCG GCC ACC GGC GGC GTG AAG AAG CCG CAC  120
Mp2.3         ... ... ... ... ..G ... ... ... ... ... ... ... ..G ... ... ... ... ... ... ...  120
Hamster       ... ... ... ... ... ... ... ... ... ... ... ... ..G ... ... ... ... ... ... ...  120
Mm53          ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ..T ...  120
Mm221-1       ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ..T ...  120
Mm221-2       ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ..A ..T ...  120
Mm291         ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ..T ...  120
Mp1.2         ... ... ... ... ... ... ..T ... ... ... ... ... ..G ... ... ... ... ... ..C ...  120
Mp1.5         ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ..C ...  120
Mp1.10        ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ..C ...  120

Mm614         CGC TAC CGG CCC GGC ACC GTG GCG CTG CGG GAG ATC CGG CGC TAC CAG AAG TCG ACC GAG  180
Mp2.3         ... ... ... ... ... ... ... ... ... ..C ... ... ... ... ... ... ... ... ... ...  180
Hamster       ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...  180
Mm53          ... ... ..T ... ... ... ... ... ... ..C ... ... ... ... ... ... ... ... ... ...  180
Mm221-1       ... ... ..T ... ... ..T ... ..A ... ..C ... ... ... ... ... ... ... ... ... ...  180
Mm221-2       ... ... ..T ... ... ... ... ..T ... ..C ... ... ... ... ... ... ... ... ... ...  180
Mm291         ... ... ..T ... ... ... ... ..C ... ..C ... ... ... ... ... ... ... ... ... ...  180
Mp1.2         ... ... ..T ... ... ... ... ... ... ..C ... ... ... ... ... ... ... ... ... ...  180
Mp1.5         ... ... ..T ... ... ... ... ... ... ..C ... ... ... ... ... ... ... ... ... ...  180
Mp1.10        ... ... ..T ... ... ... ... ... ... ..C ... ... ..A ... ... ... ... ... ... ...  180
```

FIGURE 1.—Nucleotide sequences of 10 rodent replication-dependent H3 histone genes and surrounding non-coding DNA. *M. musculus* sequences Mm614, Mm221-1, Mm221-2 and Mm291 are from TAYLOR *et al.* (1986); *M. musculus* sequence Mm53 is from GRUBER *et al.* (1990). The hamster sequence and the *M. pahari* sequences Mp1.2, Mp1.5, Mp1.10 and Mp2.3 are presented here for the first time. In the 5′-flanking sequences, putative CCAAT and TATAA box sequences are underlined. In the 3′-flanking sequences the terminal stem-loop and U7 snRNA-binding sites are underlined.

genes should be imbedded within very GC-rich flanking sequences. To account for the observation that silent sites of most mammalian genes are more GC-rich than their flanking sequences and introns, we examined five GC-rich control genes and compare their flanking sequence composition to that of the H3 genes.

By combining the relationship between 3rd codon position and intron G + C content given by D'ONOFRIO *et al.* (1991) with the relationship between intron and 5′-flanking sequence G + C content given by AÏSSANI *et al.* (1991), we can estimate the relationship between silent site and flanking sequence G + C in a typical

```
Mm614     CTG CTG ATC CGC AAG CTG CCG TTC CAG CGC TTG GTG CGC GAG ATC GCG CAG GAC TTC AAG  240
Mp2.3     ... ... ... ... ... ... ... ... ... ... C.. ... ... ... ... ... ... ... ... ...  240
Hamster   ... ... ... ... ... ... ... ... ... ... C.. ... ... ... ... ... ... ... ... ...  240
Mm53      ... ... ... ... ... ... ... ... ... ... C.. ... ... ... ... ... ... ... ... ...  240
Mm221-1   ... ... ... ... ... ... ... ... ... ... C.. ... ... ... ... ... ... ... ... ...  240
Mm221-2   ... ... ... ... ... ... ... ... ... ... C.. ... ... ... ... ... ... ... ... ...  240
Mm291     ... ... ... ... ... ... ... ... ... ... C.. ... ... ... ... ... ... ... ... ...  240
Mp1.2     ... ... ... ... ... ... ... ... ... ..G C.. ... ... ... ... ... ... ... ... ...  240
Mp1.5     ... ... ... ... ... ... ... ... ... ..G C.A ... ... ... ... ... ... ... ... ...  240
Mp1.10    ... ... ... ... ... ... ... ... ... ... C.. ... ... ... ... ..A ... ... ... ...  240

Mm614     ACG GAC CTG CGC TTC CAG AGC TCG GCC GTC ATG GCG CTG CAG GAG GCG AGC GAG GCC TAC  300
Mp2.3     ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...  300
Hamster   ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...  300
Mm53      ..C ... ... ... ... ... ... ... ..T ... ... ..T ... ... ... ... ... ... ... ...  300
Mm221-1   ..C ... ... ... ... ... ... ... ... ... ... ..T ... ... ... ..C T.T ... ... ...  300
Mm221-2   ..C ... ... ... ... ... ... ... ... ... ... ..T ... ... ... ... ... ... ... ...  300
Mm291     ..C ... ... ... ... ... ... ... ... ... ... ..T ... ... ... ..C T.T ... ... ...  300
Mp1.2     ..C ... ... ... ... ... ... ... ... ... ... ..T ... ... ... ... ... ... ... ...  300
Mp1.5     ..C ... ... ... ... ... ... ... ... ... ... ..T ... ... ... ..C T.T ... ... ...  300
Mp1.10    ..C ... ... ... ... ... ... ... ... ... ... ..T ... ... ... ..C T.T ... ... ...  300

Mm614     CTG GTG GGG CTG TTC GAG GAC ACC AAC CTG TGC GCC ATC CAC GCC AAA CGC GTC ACC ATC  360
Mp2.3     ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ..G ... ... ... ...  360
Hamster   ... ... ..T ... ... ... ... ... ... ... ... ... ... ... ... ..G ... ... ... ...  360
Mm53      ..T ... ..T ... ..T ... ... ... ... ... ... ... ... ... ... ..G ..T ... ... ...  360
Mm221-1   ..C ... ..T ... ..T ... ... ... ... ... ... ... ... ... ... ..G ..T ... ... ...  360
Mm221-2   ..T ... ..T ... ..T ... ... ... ... ... ... ... ... ... ... ..G ..T ... ... ...  360
Mm291     ..C ... ..T ... ..T ... ... ... ... ... ... ... ... ... ... ..G ..T ... ... ...  360
Mp1.2     ..T ... ..T ... ..T ... ... ... ... ... ... ... ... ... ... ..G ..T ... ... ...  360
Mp1.5     ..T ... ..T ... ..T ... ... ... ... ... ... ... ... ... ... ..G ..T ... ... ...  360
Mp1.10    ..T ... ..T ... ..T ... ... ... ... ... ... ... ... ... ... ..G ..T ... ... ...  360

Mm614     ATG CCC AAG GAC ATC CAG TTG GCC CGC CGC ATC CGT GGG GAG CGC GCT TAA GCGCCCTGTCT  422
Mp2.3     ... ... ... ... ... ... ... ... ... ... ... ... ..T ... ... ... .G. GCGCCCCGTCT  422
Hamster   ... ... ... ... ... ... ... ... ..T ... ... ..C ... ... ... ... .G. GCTCTCTATCT  422
Mm53      ... ... ... ... ... ... C.C ... ... ... ..T ... ... ... A.G ... ... ACGAAACAAAG  422
Mm221-1   ... ... ... ... ... ... C.. ... ..T ... ... ..C ... ... A.G ... ... GGGTTTCTGTT  422
Mm221-2   ... ... ... ... ... ... C.. ... ... ..T ... ..C ..C ... ..G ... ... TAGGCACGCTT  422
Mm291     ... ... ... ... ... ... C.. ... ... ... ... ..C ... ... A.G ... ... AGTTCGTCTTT  422
Mp1.2     ... ... ... ... ... ... C.. ... ... ... ... ... ..C ... A.G ... ... ACACAAGTCCA  422
Mp1.5     ... ... ... ... ... ... C.. ... ... ... ... ..C ... ... A.G ... ... GTTCGTCTTTC  422
Mp1.10    ... ..T ... ... ... ... C.. ... ... ... ... ..C ... ... A.G ..G ... ATTAACTATAC  422

Mm614     CCCATCCATTCCCCACAAAGGCTCTTTTCAGAGCCACCACAGCTTCCCGGAAAGAGCTGTTTACTTCGCTTAGAGCTTT  500
Mp2.3     CCCTTCCATCCCCCACAAAGGCTCTTTTCAGAGCCACCACTGCTTCCCCGGAAGAGCTTAACGCTTTGTCCGTACATCA  500
Hamster   TCTCCCTGTCCCCCACAAAGGCTCTTTTCAGAGCCACACTACACCTTCCCGGAAAGAGCTGTTTGACTTTAGTTTCCTC  500
Mm53      TGCAAACTGGAACCCAAAGGCTCTTTTCAGAGCCACCCACAGTTTCTGTAAAAGCAGCTGTTACTTTTCTACTGTCTTG  500
Mm221-1   AATCCACACAACCACTTTAAAGGCTCTTCTTAGAGCCACCCATCTTCCAAAAAAAAGAACTGTGCGCTTTTTCCAACTT  500
Mm221-2   TCTACACTGGCACGTAAACCAAAACGGCTCTTTTAAGAGCCACCTCCATTATCCACCAAAGATGCTTGAAGTACAAGTT  500
Mm291     CTGTGTTTTTCAAAGGCTCTTTTCAGAGCCACCCAAACCTACACAAAGGTGCTGTTGACTTATAGGCCAATTGGAGTTT  500
Mp1.2     AACTGCAACCCAAAGGCTCTTTTCAGAGCCACCCATAATTTCTGTAGAAGCAGCTGTTACATTTCTACACTCTTGGGAA  500
Mp1.5     TGTGTTTTTCAAAGGCTCTTTTCAGAGCCACCCATACCTACACAAAGGTGCTTGTCTTACTTGGTAGGGCCGATTGGAG  500
Mp1.10    AGTTCTAAGTGGTTAACCAAAGGCTCTTTTCAGAGCCACACACGACTTCAAGTAGAATTGCTGTAATAATTGATCTATT  500

Mm614     GTCCGCACATTATCCCGCTCCATAGCTCTAGGTTGGTATCTGACTTATACCTGTTATGTCTACATAATGGTTGCTTTCA  578
Mp2.3     GCCTTCTAGATAGTTCTAGGTTGGTATTTGACCTCTATTGTACCTGTTTGTCTGCATAATGTTTGCTTCCAAGGAAGCT  578
Hamster   TGCTCCCCTGTGAGTTACTGTGTAATGAAACTTTTTCTCATGAACACCTTCCTAGTAAAGTTGCAGGATCC          570
Mm53      GAAACTTTGTTTTTCAAAGATTTATCTAGTTACAAAGTATAC                                       541
Mm221-1   GTGGGTATTAATCAGTTTCATTTGTGAAAGTGCTAGGTCTCCTTTCATCACACCGATCTATTATAAGCAAGCAATATTT  578
Mm221-2   GTGAGAGTTTTCTAGGGTTTCCTATTATAGCCTTTCTTGACAATGTGAGCACCCCCCGACGAAGCAGTCTGAGTTAACA  578
Mm291     TTTCTTGGAACGGTAGCCAGGTTCTAACAGTTTTTCTTAGCTGAGGGATGCTCTTTAGAGTGTCTAAGTTGGTGAGTAT  578
Mp1.2     CTTTGTTTTTTAAAGGTTTTGTTTACTAATACAT                                               533
Mp1.5     TTCTAGGTTTGGGACGGGNAGCCAGCTTCTGAGTTTCAATTNACCTGCTTAATTTTAGCTTATGGGTGCCCTTTAAGAG  578
Mp1.10    CCTTCCTCAGGTGAACCTATATATCCCCCAGTAACACTTTAAGATGCAT                                548

Mm614     AGGAAGGTTTGAATGCCCTTAGTGTAAGTTGTTTAACGTTAACTGCGGTCTTATTCAGACGCCTAGCTTGTTAACCCGA  656
Mp2.3     T                                                                                579
Mm221-1   TGGTAAGTGATATCTTTACCCTTTTTCATGGAAATTGTACTCTTAGTTAGCAAAGCATAAAGATTGGAAAACTTTGCAT  656
Mm221-2   AAGAGGAGCTAAGCTATGCAACAAACCAGATTTCTATTGGTCACAAATTTGAAGTTGAGACCTGTTATCCAATTACCAA  656
Mm291     TCTGCAAAA                                                                        597
Mp1.5     TNCCTAAGCTGGANAGTTACTTTGGAANTGTCCAAGNANGGGAATTAATGNTTTACCTTGCC                   639

Mm614     AGTAGGTTCACCAGCTCAGTCTAAGGGGAGGGGATAATCTTACGGAGGGTTGGGGCTTGGTGGATACTATGTAATGTTC  734
Mm221-1   GGTTACAGTAATTTGGGCATTTACCGAATTTAAAAAATAAAATTATTAGTTGCAGATGTTTACTTTGGAATTGTGAACT  734
Mm221-2   GTACTTCCGCATACATCATCATAGGCATTTGAAGATTTCAACCAATCAGGAGCATGTTCCTTCTATAAAGGAACCCAGA  734

Mm614     CTACCCAGCTGGAGCAGAGAAGGC  758
Mm221-1   TTCTAAGGACAGAACCACTAATAT  758
Mm221-2   ACCTAACCTCTGCATTTCCTATTT  758
```

FIGURE 1.—Continued

A

```
Mp2.3     CGGACCCGCCCTCTACGGGCAGCTGCCAGACGTGGCTCCACCCTCGCAGAAGGTTGGCG -241
Mm614     G..................................................C......A -241

Mp2.3    ATCGGCCGGGCGCCAGGCTGGACCTGGGGGGCG---GGGGGCGGGTCTGGAGGCGTGACG -181
Mm614    .GT.A.......GG........G.........GGT...........A..T......G.. -181

Mp2.3    GGCCCGAGCCAATGGGCGAGGTCGGTGCCTGTGACGTCACGGCCAATGGCGCGGCAGCGC -121
Mm614    .........................G...G......C....................... -121

Mp2.3    GGGAGTTTCAAGTCGCTGTCTCCGCCCCGCCGCGGGGAAGACTGAGCCTATAAAGGCGGC  -61
Mm614    .........................................C...............   -61
Hamster                           GGTGG.G...........TT.........A...T  -61

Mp2.3    CGGCTCGGGCCGGTGTCAGGTCCCCGTGTGCTCCTCGTCTGGGTGTCTTCCGTCTCCGCC   -1
Mm614    ..............A...........A...........T....C...........T....   -1
Hamster  .....GC.A..T.G...TA....................T....A.............   -1
```

B

```
Mp2.3    GCGCCCCGTCTCCCTTCCATCCCCCACAAAGGCTCTTTTCAGAGC--CACCACTGCTTCC  469
Mm614    ......T.......A.....T..........................--......A......  469
Hamster  ..T.T.TA...T.TCC.TG..........................CA...T..AC.....  471

Mp2.3    CCGGAAGA------------GCTTAACGCTTTGTCCGTACATCAGCCTTCTAGATAGTTC  517
Mm614    .G.A....GCTGTTTACTTC.....GA..........C....T.T..CG..CC....C..  529
Hamster  .G.A....------------...-----G....A.TT..GT.TCCT..G..CC---CC.G  511

Mp2.3    TAGGTTGGTATTTGACCTCTATTGTACCTGTT-TGTCTGCATAATGTTTGCTTCCAAGGA  576
Mm614    ...........C....-----..A........A.....A.......G......T......  584
Hamster  .GA...AC.G.G.A.-------..A.A..T..---...-...G.--ACAC.....T..T.  558

Mp2.3    AGCTT            581
Mm614    ..G..TGAATGCCC   598
Hamster  .AG..GCAGGATCC   572
```

FIGURE 2.—Alignment of flanking sequences illustrating the orthologous relationships among the chromosome *3* H3 genes. Alignments were produced by the Pileup program in the GCG computer package using a gap weight of 3.0 and a gap length weight of 0.3, although similar alignments are found with a range of weights. (A) 5′-Flanking and nontranscribed sequences from Mp2.3, Mm614 and the hamster gene. (B) 3′-Flanking and nontranscribed sequences from Mp2.3, Mm614 and the hamster gene.

TABLE 1

Pairwise percent similarities at silent codon positions

	Chromosome *3*			Chromosome *13*						
	Mm614	Mp2.3	Hamster	Mm53	Mm221-1	Mm221-2	Mm291	Mp1.2	Mp1.5	Mp1.10
Mm614	—	95.7	96.3	85.9	85.3	86.5	87.1	86.5	85.3	84.7
Mp2.3		—	95.7	85.3	84.7	86.5	86.5	88.3	85.3	84.0
Hamster			—	85.9	87.7	87.7	88.3	87.7	86.5	85.9
Mm53				—	91.4	92.0	93.3	92.0	91.4	90.2
Mm221-1					—	92.6	97.5	90.8	93.9	92.6
Mm221-2						—	94.5	92.6	92.0	90.8
Mm291							—	92.6	95.7	94.5
Mp1.2								—	94.5	92.0
Mp1.5									—	93.9
Mp1.10										—

mammalian gene as:

$$\text{Flanking} = 9.40 + 0.5957 \times \text{Silent}.$$

Selection, but not mutation bias, should be strongest on highly expressed genes. Our controls include genes with very high third position G + C content that are expressed at lower levels than the H3 genes. Only a small fraction of mammalian genes have silent site G + C content as high as that seen in the H3 histone genes. Of 363 rat and mouse genes compared by WOLFE and SHARP (1993), only three show third position G + C content of 85% or higher. Interestingly, all three of these genes are transcription factors (AGP/EGP, C/EBP, and SCIP). We also examined human transforming protein hst

TABLE 2

Scaled-χ^2 values

Genes	Scaled-χ^2
Chromosome *3*	
Mm614	1.556
Mp2.3	1.625
Hamster	1.616
Chromosome *13*	
Mm53	1.299
Mm221-1	1.488
Mm221-2	1.504
Mm291	1.616
Mp1.2	1.433
Mp1.5	1.418
Mp1.10	1.588

TABLE 3

G + C content of H3 and control genes

Gene	Percent GC Third position	Percent GC Predicted 5′	Percent GC (*n*) 5′ UTR	Percent GC (*n*) Total 5′
C/EBP (rat)	86.0	60.6	73.5 (132)	60.2 (465)
AGP/EBP (rat)	95.2	66.1	83.0 (53)	83.0 (119)
SCIP (mouse)	85.4	60.3	97.1 (35)	
HST (human)	93.7	65.2	78.2 (238)	63.8 (2550)
XIHB (human)	94.4	65.6	72.4 (29)	
Mm614	97.0	67.2	65.1 (43)	65.3 (803)
Mm53	86.5	60.9		48.9 (137)
Mm221-1	88.0	61.8	46.4 (28)	58.3 (273)
Mm221-2	88.0	61.8	23.8 (21)	37.4 (275)
Mm291	91.7	64.0	48.4 (31)	39.9 (293)
Mp1.2	89.5	62.7		37.8 (45)
Mp1.5	88.7	62.2		39.9 (293)
Mp1.10	90.2	63.1		43.0 (286)

(HUMHST) and human ζ-globin (HUMXIHB), both of which have third position G + C content over 90%.

For all five control genes, the observed 5′-flanking sequence G + C content is as high as or higher than the prediction (Table 3). In contrast, the combined 5′-flanking sequences for the chromosome *13* H3 genes are approximately 43.9% G + C (after first eliminating the known regulatory and structural elements that are underlined in Figure 1), well below the 62.4% G + C predicted by the third position composition. However, the G + C content of 5′-flanking sequences of the chromosome *3* gene (65.3%) is very close to the predicted value of 67.2%.

The Mm614 5′-flanking sequence may not accurately reflect the overall G + C content of the region of chromosome *3* on which this gene is located. The 5′-flanking sequences are the intergenic region between the upstream H2A gene and the H3 gene (HURT *et al.* 1989). At least some of those sequences have a regulatory function, so this region may be under selection. The Mm614 3′-flanking sequence beyond the U7-binding site is significantly lower in G + C content compared to the 5′-flanking sequences, with only 46.1% G + C ($\chi^2 = 52.6$, $P < 0.001$). Additionally, the 5′-promoter region of the upstream H2A gene on chromosome *3* is also GC-rich, but only for about 250 bp. Beyond that distance the G + C content again drops to 44.1% (HURT *et al.* 1989). Thus, the chromosome *3* H3 gene is embedded within a GC-rich region, but that region includes only about 2 kb and is composed mostly of coding sequences and known or suspected regulatory sequences. This 2 kb of GC-rich sequence apparently is itself embedded within an AT-rich region, so it is not clear if we should expect mutation to be strongly biased toward G and C or weakly biased toward A and T in the chromosome *3* gene. If the high G + C content of the Mm614 gene is caused by a strong mutational bias, then the sharp transition from an A + T bias to a G + C bias and back again within 2 kb is itself a noteworthy phenomenon.

Dinucleotide mutation bias effects: It is possible that mutation pressure on the histone coding sequences is not simply a reflection of the base frequencies in surrounding noncoding DNA. Mutation frequencies are known to be affected by base context (BULMER 1986), so the bias seen in silent sites may be due to the adjoining replacement sites. In that case the codon bias could be due to constraints imposed by the H3 amino acid sequence rather than to selection on silent sites. If a dinucleotide bias does exist, it may be difficult to detect with statistical analyses in the H3 genes, because of their small size. Clearly, it is not appropriate to pool all of the individual genes to increase the sample sizes, because it is not clear how recently these genes may have shared a common ancestor due to gene conversion. It does appear that there has not been any gene conversion between chromosome *3* and chromosome *13* (see below), so it may be appropriate to pool one gene from each chromosome. When the sample size from a single gene is too small, we have chosen to combine counts from two *M. pahari* genes, Mp2.3 from chromosome *3* and Mp1.2 from chromosome *13* (Table 4).

It is unlikely that dinucleotide-based mutation bias would always favor NG and NC dinucleotides, so some evidence against dinucleotide effects is provided by the observation that codons ending in A or T are never preferred. Relative preference for C or G can only be compared among fourfold degenerate codons. We first examine the possible influence of the first base of the following codon (the "fourth" position of the codon) on the choice between G and C in the third position. To hold constant any influence of the preceding base, we consider only those codons with a C in the second position (EYRE-WALKER 1991). We find no evidence that the choice between G and C in the third position is influenced by the nucleotide at the fourth position (Table 5).

We next consider whether the nucleotide preceding a fourfold silent site influences the frequencies of G and C at the silent sites. In this case there are enough codons that we can examine only the chromosome *3* gene Mm614. We find a significant effect (Table 6), with an excess of G following a second position T and an excess of C following a second position G. If this effect of the second position on the third is real, then it is surprising that we do not also see an effect of the fourth position on the third position. In particular, there is no evidence for an avoidance of GG dinucleotides in the third position to fourth position comparison (Table 5).

Use of C or G in the third position might instead be due to an effect of the first position on the mutation pattern at the third. However, the pattern observed for the Leu and Arg codons rejects this possibility. Both groups of codons have C in the first position, yet they differ significantly in use of C or G in the third position ($\chi^2 = 17.64$, $P < 0.001$, using only the Mp2.3 gene).

Comparisons based on the nucleotide in the second position will necessarily confound dinucleotide-based

TABLE 4

Codon usage in two *M. pahari* H3 histone genes

		Mp2.3	Mp1.2			Mp2.3	Mp1.2			Mp2.3	Mp1.2			Mp2.3	Mp1.2
TTG	Leu	1	0	TCG	Ser	2	2	TAG	End	0	0	TGG	Trp	0	0
TTA	Leu	0	0	TCA	Ser	0	0	TAA	End	0	1	TGA	End	1	0
TTT	Phe	0	1	TCT	Ser	0	0	TAT	Tyr	0	0	TGT	Cys	0	0
TTC	Phe	4	3	TCC	Ser	1	1	TAC	Tyr	3	3	TGC	Cys	1	1
CTG	Leu	11	11	CCG	Pro	4	3	CAG	Gln	8	8	CGG	Arg	2	2
CTA	Leu	0	0	CCA	Pro	0	0	CAA	Gln	0	0	CGA	Arg	0	0
CTT	Leu	0	1	CCT	Pro	0	0	CAT	His	0	0	CGT	Arg	2	5
CTC	Leu	0	0	CCC	Pro	2	3	CAC	His	2	2	CGC	Arg	14	10
ATG	Met	3	3	ACG	Thr	3	1	AAG	Lys	13	13	AGG	Arg	0	1
ATA	Ile	0	0	ACA	Thr	0	1	AAA	Lys	0	0	AGA	Arg	0	0
ATT	Ile	0	0	ACT	Thr	0	0	AAT	Asn	0	0	AGT	Ser	0	0
ATC	Ile	7	7	ACC	Thr	7	8	AAC	Asn	1	1	AGC	Ser	3	3
GTG	Val	4	4	GCG	Ala	6	3	GAG	Glu	7	7	GGG	Gly	1	0
GTA	Val	0	0	GCA	Ala	0	0	GAA	Glu	0	0	GGA	Gly	0	0
GTT	Val	0	0	GCT	Ala	1	5	GAT	Asp	0	0	GGT	Gly	1	1
GTC	Val	2	2	GCC	Ala	11	10	GAC	Asp	4	4	GGC	Gly	5	6

TABLE 5

Dinucleotide frequencies for silent sites and following first position sites for codons with C in the second position

Silent nucleotide	Following nucleotide	Observed	Expected[a]
G	G	7	5.9
C	G	12	13.1
G	C	7	5.9
C	C	12	13.1
G	T[b]	2	
C	T[b]	2	
G	A	5	7.2
C	A	18	15.8

Counts for the chromosome *3* gene Mp2.3 and the chromosome *13* gene Mp1.2 are combined.

[a] Overall $\chi^2 = 1.52$ (NS).

[b] Silent bases preceding a T and silent bases other than G or C are ignored.

TABLE 6

Dinucleotide frequencies for second position sites and silent sites at fourfold degenerate codons in gene Mp2.3

Second position nucleotide	Silent nucleotide[a]	Observed	Expected[b]
G	G	3	9.7
G	C	19	12.3
C	G	15	15.8
C	C	21	20.2
T	G	15	7.5
T	C	2	9.5

[a] Silent bases other than G or C are ignored.

[b] Overall $\chi^2 = 21.8$ ($P < 0.001$).

mutation effects with selection effects that are due to a particular amino acid. In the present comparison the largest contributions to the χ^2 value come from the exclusive use of CTG codons for Leu and the strong preference for CGC codons for Arg. We do not know if these preferences are due to selection or to the presence of the preceding T or G. However, we also find a preference for CTG and CGC codons in the five control gene sequences, and the pattern in those genes is statistically indistinguishable from the pattern in the Mp2.3 gene (results not shown). If we assume that the control genes are expressed at a low enough level that selection cannot operate on synonymous codons, then we must conclude that the preference for CTG over CTC and CGC over CGG is determined in large part by context-dependent mutational effects (or by a similar alternative, such as biased repair).

Comparative analyses

Phylogenetic relationships among the H3 coding sequences: In addition to selection and mutation bias, codon usage in the replication-dependent H3 gene family could be influenced by gene conversion. Frequent gene conversion between different gene copies can result in genes evolving in concert within a species (ZIMMER *et al.* 1980). In the absence of gene conversion and prior to the silent sites becoming saturated for substitutions, each H3 gene will be most closely related to its orthologous copy in the other species. If frequent gene conversion has resulted in a complete homogenization of the H3 gene family since the divergence of *M. musculus* and *M. pahari*, then all of the *M. musculus* coding sequences will be descended from a single ancestral gene that will be different from the ancestral *M. pahari* H3 gene. If gene conversion occurs only between genes on the same chromosome, then the chromosome *3* genes from *M. musculus*, *M. pahari* and hamster will form a group that is separate from the chromosome *13* genes.

Gene conversion between chromosome *3* and chromosome *13*: Historical relationships among the coding sequences were explored by parsimony analysis using PAUP 3.1 (SWOFFORD 1993). All of the variable positions are silent, except for the single replacement substitution that distinguishes the H3.1 protein subtype from the H3.2 subtype. When only the silent positions are included, the chromosome *3* genes from all three species

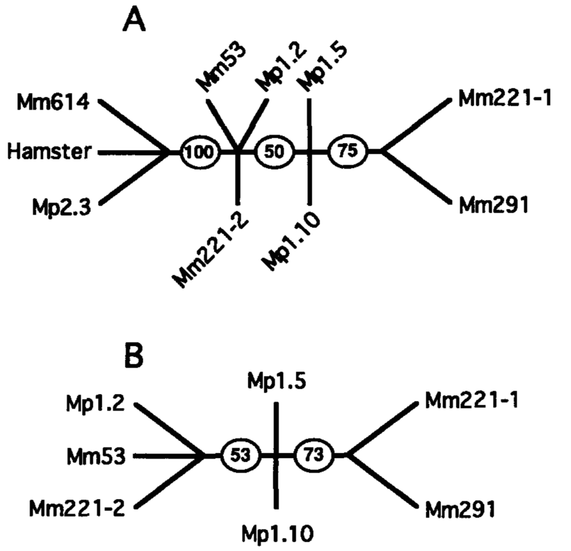

FIGURE 3.—Parsimony analysis of the historical relationships among the H3 genes. Trees are unrooted. Circled numbers refer to the proportion of trees resulting from 1000 bootstrap replicates using a branch and bound search algorithm that included the monophyletic groups defined by that bipartition. Bipartitions not found in at least 50% of the bootstrap replicates are collapsed into multichotomies. Only silent sites are included in these analyses. (A) Analysis of all 10 H3 gene sequences. (B) Analysis of only the chromosome *13* genes.

cluster separately from any of the chromosome *13* genes on the most parsimonious tree. Examination of the informative characters shows that there are 12 sites at which the chromosome *3* genes all have a character state that is shared by none of the chromosome *13* genes. This large distinction is also strongly supported by a bootstrap analysis (FELSENSTEIN 1985)–the chromosome *3* genes were monophyletic in 100% of 1000 bootstrap replicates (Figure 3*A*). This phylogenetic analysis suggests that there have been no gene conversions from chromosome *13* to chromosome *3* since the divergence between cricetid rodents (represented by the hamster) and murid rodents. It is also appears that there have been no or very few conversions from chromosome *3* to chromosome *13*. One or more of the chromosome *13* genes that have not been examined may have been converted by the chromosome *3* gene, but no chromosome *3*-specific nucleotides have spread through the chromosome *13* array.

After a gene conversion between the two chromosomes, the genes involved will begin to diverge. It is possible that the large distinction between genes from the two chromosomes is due to the silent sites becoming saturated for substitutions quickly, relative to the time between gene conversions. If so, then the genes repeatedly return to a chromosome-specific codon usage pattern. Thus, the phylogenetic analysis shows that gene conversions between chromosome *3* and chromosome *13* are either rare or have only a transitory effect. The evolutionary independence of the single chromosome *3* gene from the cluster on chromosome *13* means that there is, in effect, both a replication-dependent H3 multigene family on chromosome *13* and a separate replication-dependent H3 single copy gene on chromosome *3*.

Gene conversion within the chromosome *13* array: The chromosome *13* genes are expressed at lower levels than the chromosome *3* gene (GRAVES *et al.* 1985). This should cause the chromosome *13* genes to be less affected by selection, yet they have silent site G + C contents that are much higher than expected based on the flanking sequence G + C composition. It is possible that the high level of homogeneity among the chromosome *13* genes may be due at least in part to frequent gene conversion rather than selection.

Unfortunately, the phylogenetic relationships among the chromosome *13* genes cannot be clearly resolved. No grouping of chromosome *13* genes is strongly supported by bootstrap analysis, whether the chromosome *3* genes are included or not (Figure 3). This is primarily due to the very high similarity among all of the chromosome *13* genes. Among the seven chromosome *13* genes there are only 10 differences that are informative for parsimony analysis. The lack of phylogenetic resolution provides some evidence that gene conversion has not been frequent enough to cause complete turnover within the chromosome *13* genes since the divergence between *M. musculus* and *M. pahari*. Such a process of concerted evolution should produce species-specific sites, but there is only a single site where the *M. musculus* genes all share one state while the *M. pahari* genes all share a different state.

Again, it is possible that the chromosome *13* genes quickly become saturated for silent substitutions following gene conversions, thus obscuring any phylogenetic evidence for concerted evolution via gene conversion. However, differences between the chromosome *3* genes and the chromosome *13* genes do not appear to be randomly distributed, as would be expected if these genes were all saturated for silent substitutions. On average, each chromosome *13* gene is different from a chromosome *3* gene at 21.8 silent sites, but the chromosome *13* genes average only 3.6 sites each where a gene has a unique silent site. At 10 fourfold degenerate sites the three chromosome *3* genes all share a state found in none of the chromosome *13* genes. At 7 of those 10 sites all of the chromosome *13* genes have the same alternative state. It is very unlikely that these seven substitutions would have occurred independently in all seven chromosome *13* genes. It is also difficult to imagine a scenario where one silent state would be favored by selection at a particular position in the chromosome *3* gene while a different state was favored for the chromosome *13* genes. Thus, some of the homogeneity among the chromosome *13* genes appears to be due to gene conversion. Conversion is frequent enough to result in concerted evolution between the chromosomes, but not frequent enough to result in concerted evolution on chromosome *13* between the two species of Mus.

Examination of the pair of orthologous chromosome *13* genes (Mm291 and Mp1.5) could also provide evidence regarding gene conversion. Without gene conversion, these two coding sequences should be closest relatives. That relationship is not supported by the bootstrap analysis, and in fact a monophyletic group containing only Mm291 and Mp1.5 did not appear on any of the most parsimonious trees in any of the bootstrap replicates. If the Mm291 and Mp1.5 coding regions are each other's closest relatives we would expect Mm291 and Mp1.5 to share some nucleotides at silent sites that are not found in any other chromosome *13* genes. However, these two genes do not uniquely share any silent states. Mp1.5 only shares rare silent states with the other *M. pahari* genes; one site is shared with Mp1.10 and two sites are shared with Mp1.2. Mm291 shares one rare state, and that is with another *M. musculus* gene, Mm221-1. Overall, Mm291 and Mm221-1 are the most similar pair of genes among all those examined, differing at only 4 positions (Table 1). These data suggest the possibility that Mm291 has been converted by either Mm221-1 or by another, very similar *M. musculus* gene.

DISCUSSION

Biased usage of synonymous codons certainly can be generated without natural selection. High G + C content will necessarily produce biased codon usage, and G + C content bias can be caused by biased mutational mechanisms, either differential misincorporation, differential repair, or biased gene conversion during recombination (EYRE-WALKER 1993). It is clear that in the large majority of mammalian genes the pattern of codon usage is consistent with bias resulting from mutation pressure; there is no need to invoke selection on synonymous codon usage (WOLFE *et al.* 1989; SHARP 1989; EYRE-WALKER 1991; WOLFE and SHARP 1993).

The action of natural selection is always difficult to definitively prove. However, the pattern of codon usage, particularly compared to the pattern of flanking sequence nucleotide usage, strongly suggests that mutational effects are not entirely responsible for the observed codon-usage pattern. The relationship between silent site base composition and flanking sequence base composition is very different in the H3 genes compared to the poorly expressed control genes. Silent site base composition in the H3 genes is substantially higher than predicted based on the typical pattern in mammalian genes. If the high G + C content in the H3 silent sites is due to mutation pressure rather than selection, then the mutational mechanism must not be based on transcription, as it only affects the 411 nucleotide region that is translated into protein.

Dinucleotide effects appear to have only a secondary influence on codon usage in the H3 genes. Silent sites are independent of the state at the following position. The preceding position exerts some influence, but only on the preference for C or G at some fourfold degenerate sites. Within the chromosome *13* array, it is possible that some of the homogeneity among genes can be attributed to gene conversion between genes. However, gene conversion alone would produce sequence homogeneity but not necessarily biased codon usage. Biased gene conversion within a single gene could favor high G + C (EYRE-WALKER 1993). However, when two copies of the same gene form a heteroduplex there is no barrier to prevent the heteroduplex from extending into the flanking regions. Again, any bias in heteroduplex repair would have to operate only on sequences that are destined to be translated.

When selection does act on codon usage, selection coefficients associated with variation at silent sites are assumed to be very small (LI 1987), so silent site variation should be effectively neutral except in very large populations. It is not clear what the effective population size might have been for *M. musculus* prior to the establishment of its commensal relationship with human populations, but populations of *M. pahari* and hamsters clearly are several orders of magnitude smaller than those of *E. coli*, and are likely one to three orders of magnitude smaller than those of most Drosophila. With these relatively small effective population sizes, maintenance of any codon bias by selection alone would require much larger fitness penalties against suboptimal codons than those inferred for unicellular organisms. If any mammalian genes might be subject to codon selection strong enough to overcome the effects of drift, the replication-dependent histone genes are good candidates. These genes, particularly the chromosome *3* genes, are expressed at extremely high levels during S phase of the cell cycle in every cell in the animal (GRAVES *et al.* 1985).

Selective differences between synonymous codons, and consequently codon-usage bias, should be greatest in highly expressed genes. This is true in unicellular organisms, where a strong correlation is found between the level of expression and degree of codon bias (GOUY and GAUTIER 1982; IKEMURA 1985; SHARP *et al.* 1986; SHARP and LI 1987; SHARP and DEVINE 1989). Likewise, in Drosophila, MORIYAMA and HARTL (1993) found that codon bias in the *Adh* gene is strongest in those species where *Adh* is most highly expressed. It is not entirely clear if there is a relationship within the replication dependent H3 genes between the level of expression and degree of codon bias. The chromosome *13* genes are each expressed approximately 10-fold less than the chromosome *3* gene (GRAVES *et al.* 1985), and they do show a less pronounced codon-usage bias than the chromosome *3* genes. This interpretation would be valid if the chromosome *3* mutation pressure is about the same as that for chromosome *13* (as indicated by the sequences 5′ from the H2A gene and 3′ from the H3 gene). However, if the chromosome *3* gene is subject to higher mutation pressure toward G and C (as indicated by the intergenic region 5′ from the H3 gene), then the higher codon usage bias seen in the chromosome *3* gene might not be related to the higher level of expression.

If selection is operating on the H3 silent sites, then it is possible that gene conversion helps to maintain optimal codon usage in the chromosome *13* cluster. If a gene with several non-preferred codons converts a gene with fewer non-preferred codons, that would have the same effect as several simultaneous mutations at silent sites and the converted product would be selected against. Conversely, a conversion event where the number of non-preferred codons is decreased would give a product favored by selection. Thus, codon selection combined with gene conversion could produce an effect similar to biased gene conversion, but one that would be limited to only the coding sequence.

HUYNEN *et al.* (1992) suggested that vertebrate histone genes in general are under selection for mRNA secondary structure, although they did not propose any specific structures that might be important. Under their model, selection favors frequencies of G and C at silent sites that balance the G's and C's at all other sites, so that in the mRNA molecule the frequency of G is approximately equal to the frequency of C. This model appears more plausible than the major alternative, that favored codons are those that match the most abundant tRNA molecules (BULMER 1987b). The secondary structure hypothesis accounts for the absence of any codons where A or T are preferred, while there is no *a priori* reason to expect that abundant tRNAs would match only G- or C-ending codons. However, nothing is currently known about secondary structure requirements of histone mRNAs and the relative abundances of different tRNA species in rodents are not known. Another alternative is that selection might be acting to minimize the translational error rate (BULMER 1991). Clearly there is extremely strong selection operating on the amino acid sequence of the replication-dependent H3 protein. This selection operates over phylogenetic time, as exemplified by the remarkable similarity in H3 sequence even between animals and plants (DELANGE *et al.* 1973; PATTHY *et al.* 1973). It also probably acts across individual members of the H3 multigene family within organisms. Unlike the H2A and H2B gene families, where DNA sequence data have revealed predicted amino acid sequence variation that had not been previously observed at the protein level (LIU *et al.* 1987), we find no variation in predicted H3 protein sequence aside from the widely distributed H3.1 and H3.2 subtypes. It is possible that even small amounts of H3 protein with incorrect peptide sequence could be detrimental to the organism.

We thank S. RUDIKOFF for providing the *M. pahari* genomic library, and A. LEE for providing plasmid pAAD3.7 containing the hamster chromosome *3* H3 gene. We thank SCOTT WILLIAMS for providing some of the original impetus for this project, and for numerous discussions. We also thank two anonymous referees for comments that resulted in distinct improvements in this manuscript. DNA sequences were determined at the University of North Carolina, Chapel Hill automated sequencing facility using the ABI 373A and the *Taq* Dyedeoxy terminator cycle sequencing kit, and by ELLYN WHITEHOUSE at the Florida State University automated sequencing facility using the ABI 373A. This work was supported by National Institutes of Health postdoctoral fellowship GM13200 (R.W.D.) and grant GM29832 (W.F.M.). Completion of this work was partially supported by an Alfred P. Sloan Foundation Fellowship in Studies of Molecular Evolution (R.W.D.).

LITERATURE CITED

AÏSSANI, B., G. D'ONOFRIO, D. MOUCHIROUD, K. GARDINER, C. GAUTIER *et al.*, 1991 The compositional properties of human genes. J. Mol. Evol. **32:** 493–503.

AOTA, S., and T. IKEMURA, 1986 Diversity of G+C content at the third positions of codons in vertebrate genes and its cause. Nucleic Acids Res. **14:** 6345–6355.

ARNHEIM, N., 1983 Concerted evolution of multigene families, pp. 38–61 in *Evolution of Genes and Proteins*, edited by M. NEI and R. K. KOEHN. Sinauer, Boston.

ARTISHEVSKY, A., S. WOODEN, A. SHARMA, E. RESENDEZ, JR. and A. S. LEE, 1987 Cell-cycle regulatory sequences in a hamster histone promoter and their interactions with cellular factors. Nature **328:** 823–827.

BERNARDI, G., B. OLOFSSON, J. FILIPSKI, M. ZERIAL, J. SALINAS *et al.*, 1985 The mosaic genome of warm-blooded vertebrates. Science **228:** 953–958.

BULMER, M., 1986 Neighbouring base effects on substitution rates in pseudogenes. Mol. Biol. Evol. **3:** 322–329.

BULMER, M., 1987a A statistical analysis of nucleotide sequences of introns and exons in human genes. Mol. Biol. Evol. **4:** 395–405.

BULMER, M., 1987b Coevolution of codon usage and transfer RNA abundance. Nature **325:** 728–730.

BULMER, M., 1991 The selection-mutation-drift theory of synonymous codon usage. Genetics **129:** 897–907.

DELANGE, R. J., D. M. FAMBROUGH, E. L. SMITH and J. BONNER, 1969 Calf and pea histone IV. III. Complete amino acid sequence of pea seedling histone IV: comparison with the homologous calf thymus histone. J. Biol. Chem. **244:** 5669–5679.

DELANGE, R. J., J. A. HOOPER and E. L. SMITH, 1973 Histone III. Sequence studies on the cyanogen bromide peptides: complete amino acid sequence of calf thymus histone III. J. Biol. Chem. **248:** 3261–3274.

D'ONOFRIO, G., D. MOUCHIROUD, B. AÏSSANI, C. GAUTIER and G. BERNARDI, 1991 Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. J. Mol. Evol. **32:** 504–510.

DOVER, G. A., 1982 Molecular drive: a cohesive mode of species evolution. Nature **299:** 111–117.

ENGEL, J. D., and J. B. DODGSON, 1981 Histone genes are clustered but not tandemly repeated in the chicken genome. Proc. Natl. Acad. Sci. USA **78:** 2856–2860.

EYRE-WALKER, A., 1991 Analysis of codon usage in mammals: selection or mutation bias? J. Mol. Evol. **33:** 442–449.

EYRE-WALKER, A., 1993 Recombination and mammalian genome evolution. Proc. R. Soc. Lond. B **252:** 237–243.

FEINBERG, A. P., and B. VOGELSTEIN, 1983 A technique for radiolabelling DNA restriction fragments to high specific activity. Anal. Biochem. **132:** 6.

FELSENSTEIN, J., 1985 Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39:** 783–791.

FILIPSKI, J., 1987 Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. FEBS Lett. **217:** 184–186.

GOUY, M., and C. GAUTIER, 1982 Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. **10:** 7055–7074.

GRAVES, R. A., S. E. WELLMAN, I.-M. CHIU and W. F. MARZLUFF, 1985 Differential expression of two clusters of mouse histone genes. J. Mol. Biol. **183:** 179–194.

GRUBER, A., A. STREIT, M. REIST, P. BENNINGER, R. BOEHNI *et al.*, 1990 Structure of a mouse histone-encoding gene cluster. Gene **95:** 303–304.

HENTSCHEL, C. C., and M. L. BIRNSTIEL, 1981 The organization and expression of histone gene families. Cell **25:** 301–313.

HILLIS, D. M., and M. T. DIXON, 1991 Ribosomal DNA: molecular evolution and phylogenetic inference. Q. Rev. Biol. **66:** 411–453.

HOOD, L., J. H. CAMPBELL and S. R. C. ELGIN, 1975 The organization, expression and evolution of antibody genes and other multigene families. Annu. Rev. Genet. **9:** 305–353.

HURT, M. M., N. CHODCHOY and W. F. MARZLUFF, 1989 The mouse histone H2a.2 gene from chromosome 3. Nucleic Acids Res. **17:** 8876.

HUYNEN, M. A., D. A. M. KONINGS and P. HOGEWEG, 1992 Equal G and C contents in histone genes indicate selection pressures on mRNA secondary structure. J. Mol. Evol. **34:** 280–291.

IKEMURA, T., 1985 Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2:** 13–34.

KLIMAN, R. M., and J. HEY, 1993 Reduced natural selection with low recombination in *Drosophila melanogaster. Mol. Biol. Evol.* **10:** 1239–1258.

LI, W.-H., 1987 Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J. Mol. Evol. **24:** 337–345.

LISKAY, R. M., A. LETSOU and J. L. STACHELEK, 1987 Homology requirement for efficient gene conversion between duplicated chromosomal sequences in mouse cells. Genetics **115:** 161–167.

LIU, T.-J., L. LIU and W. F. MARZLUFF, 1987 Mouse histone H2A and H2B genes: four functional genes and a pseudogene undergoing gene conversion with a closely linked functional gene. Nucleic Acids Res. **15:** 3023–3039.

MARZLUFF, W. F., 1986 Evolution of histone genes, pp. 139–169 in *DNA Systematics*, edited by S. K. DUTTA. CRC Press, Boca Raton, Fla.

MARZLUFF, W. F., and R. A. GRAVES, 1984 Organization and expression of mouse histone genes, pp. 281–315 in *Histone Genes: Structure, Organization and Function*, edited by G. STEIN, J. STEIN and W. F. MARZLUFF. John Wiley & Sons, New York.

MAXSON, R., T. MOHUN, G. GORMEZANA, G. CHILDS and L. H. KEDES, 1983 Distinct organizations and patterns of expression of early and late histone gene sets in the sea urchin. Nature **301:** 120–125.

MORIYAMA, E. N., and D. L. HARTL, 1993 Codon usage bias and base composition of nuclear genes in Drosophila. Genetics **134:** 847–858.

MOWRY, K. L., and J. A. STEITZ, 1987 Identification of the human U7 snRNP as one of several factors involved in the 3′ end maturation of histone pre-messenger RNAs. Science **238:** 1682–1687.

NEI, M., and D. GRAUR, 1984 Extent of protein polymorphism and the neutral mutation theory. Evol. Biol. **17:** 73–118.

OHTA, T., 1980 *Evolution and Variation in Multigene Families: Lecture Notes in Biomathematics*, Vol. 37. Springer-Verlag, New York.

PATTHY, L., E. L. SMITH and J. JOHNSON, 1973 Histone III. IV. The amino acid sequence of pea embryo histone. J. Biol. Chem. **248:** 6834–6840.

SEILER-TUYNS, A., and M. BIRNSTIEL, 1981 Structure and expression in L cells of a cloned H4 histone gene of the mouse. J. Mol. Biol. **151:** 607–625.

SHARP, P. M., 1989 Evolution at 'silent' sites in DNA, pp. 23–13 in *Evolution and Animal Breeding: Reviews on Molecular and Quantitative Approaches in Honour of Alan Robertson*, edited by W. G. HILL and T. F. C. MACKAY. Wallingford CAB International, U.K.

SHARP, P. M., and K. M. DEVINE, 1989 Codon usage and gene expression level in *Dictyostelium discoideum*: highly expressed genes do "prefer" optimal codons. Nucleic Acids Res. **17:** 5029–5039.

SHARP, P. M., and W.-H. LI, 1986 Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for "rare" codons. Nucleic Acids Res. **14:** 7737–7749.

SHARP, P. M., and W.-H. LI, 1987 The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. **4:** 222–230.

SHARP, P. M., T. M. F. TUOHY and K. R. MOSURSKI, 1986 Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. **14:** 5125–5143.

SHIELDS, D. C., P. M. SHARP, D. C. HIGGINS and F. WRIGHT, 1988 Silent sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. **5:** 704–716.

SITTMAN, D. B., R. A. GRAVES and W. F. MARZLUFF, 1983 Structure of a cluster of mouse histone genes. Nucleic Acids Res. **11:** 6679–6696.

SMITH, G. P., 1976 Evolution of repeated DNA sequences by unequal crossover. Science **191:** 528–535.

STEPHENSON, E. C., H. P. ERBA and J. G. GALL, 1971 Nucleic characterization of a cloned histone gene cluster of the newt *Notophthalmus viridescens.* Nucleic Acids Res. **9:** 2281.

SWOFFORD, D. L., 1993 PAUP: phylogenetic analysis using parsimony, version 3.0s. Computer program distributed by the Illinois Natural History Survey, Champaign.

TAYLOR, J. D., S. E. WELLMAN and W. F. MARZLUFF, 1986 Sequences of four mouse histone H3 genes: implications for evolution of mouse histone genes. J. Mol. Evol. **23:** 242–249.

WOLFE, K. H., and P. M. SHARP, 1993 Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. J. Mol. Evol. **37:** 441–456.

WOLFE, K. H., P. M. SHARP and W.-H. LI, 1989 Mutation rates differ among regions of the mammalian genome. Nature **337:** 283–285.

ZERNICK, M., N. HEINTZ, I. BOIME and R. G. ROEDER, 1980 *Xenopus laevis* histone genes: variant H1 genes are present in different clusters. Cell **22:** 807.

ZIMMER, E., S. L. MARTIN, S. M. BEVERLY, Y. W. KAN and A. C. WILSON, 1980 Rapid duplication and loss of genes for the alpha-chains of hemoglobin. Proc. Natl. Acad. Sci. USA **77:** 2158–2162.

ZWEIDLER, A., 1984 Core histone variants of the mouse: primary structure and differential expression, pp. 339–369 in *Histone Genes: Structure, Organization and Function*, edited by G. STEIN, J. STEIN and W. F. MARZLUFF. John Wiley & Sons, New York.

Communicating editor: M. BULMER