

The Compositional Evolution of the Murid Genome

Nick G.C. Smith,¹ Adam Eyre-Walker²

¹ Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, SE-752 36 Uppsala, Sweden

² Centre for the Study of Evolution & School of Biological Sciences, University of Sussex, Sussex, UK

Received: 20 July 2001 / Accepted: 25 January 2002

Abstract. Murid rodents show much less variation in isochore base composition than do most other mammals, a difference which has been referred to as the murid shift. We have investigated the murid shift by asking (1) whether the murid shift is ongoing and (2) whether there is any evidence of selection or biased gene conversion affecting base composition in the present-day mouse genome. By estimating the ancestral base composition of protein-coding genes in murids we can confirm that the murid shift is ongoing. Tests using nongenic polymorphism data fail to reject the hypothesis that base composition is due to mutation bias alone. However, the patterns of compositional change suggested by the polymorphism and divergence data differ, suggesting the possibility of two murid shifts.

Key words: Murid shift — Isochores — Selection — Base composition — Biased gene conversion — Non-equilibrium composition

Introduction

Base composition along mammalian chromosomes is homogeneous over large scales (Bernardi 2000), with the regions of similar composition referred to as “isochores.” Although all mammals appear to have isochores, not all mammals show the same isochore pattern. Most mammals share a common pattern with a moderately high level of variation in GC content between isochores. In

contrast, murids show much less variation in isochore GC content than other mammals, although their mean genomic GC content is similar (Mouchiroud et al. 1988; Robinson et al. 1997). Use of a maximum likelihood method for the inference of ancestral GC composition (Galtier and Gouy 1998) has revealed that the murid pattern is a derived state and that the nonrodent pattern is the ancestral state (Galtier and Mouchiroud 1998). This change in the genome of murids is termed the murid shift (Mouchiroud et al. 1988).

Any explanation of the murid shift will clearly be related to more general explanations for the existence of isochores. The attempt to understand isochores has led to an ongoing debate between neutralists and selectionists. The common neutralist explanation for genomic composition variation is that isochores are a consequence of variation in the pattern of mutation bias (Wolfe et al. 1989). However, recent tests have shown that mutation bias is not sufficient to explain the synonymous codon usage of mammals (Eyre-Walker 1999; Smith and Eyre-Walker 2001). Given that the GC content of protein-coding genes is highly correlated with that of the isochore in which the gene is embedded (Clay et al. 1996), these results suggest that either selection or biased gene conversion may be responsible for mammalian isochores.

We have investigated the murid shift by asking two questions. First, is the murid shift ongoing? We have inferred the evolution of base composition in murids using a large number of protein-coding genes, and we have found that compositional change is ongoing in the murids. Second, is there any evidence of selection or biased gene conversion affecting base composition in the present day murids? We have used polymorphism data from

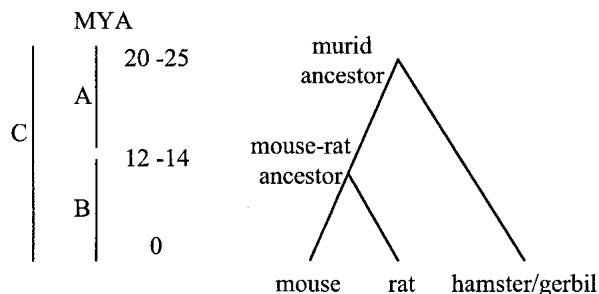


Fig. 1. Schematic phylogeny of murid evolution.

random genomic single nucleotide polymorphisms (SNPs) in the mouse to test the hypothesis that mutation bias alone is responsible for compositional variation. The polymorphism data fail to provide significant evidence against the mutation bias hypothesis.

Materials and Methods

Is the Murid Shift Ongoing?

We inferred the ancestral base composition at the third codon position of protein-coding genes using a data set of 123 genes for which we found coding sequences from mouse, rat, and one other murid (generally hamster or gerbil) and for which orthology was confirmed using Hovergen (Duret et al. 1994). Protein alignments were performed using CLUSTALW (Thompson et al. 1994), and DNA alignments were reconstructed using the program MRTRANS (written by Bill Pearson and available at www.hgmp.mrc.ac.uk).

Galtier and Gouy's (1998) method, as implemented in their NHML program, was used to estimate the ancestral GC contents at the third codon position of each gene, using the phylogeny given in Fig. 1. Our use of the NHML program revealed slight discrepancies between the results obtained using the currently available versions of NHML ([ftp://pbil.univ-lyon1.fr/pub/mol_phylogeny/nhml](http://pbil.univ-lyon1.fr/pub/mol_phylogeny/nhml)) and previously published results (Galtier and Mouchiroud 1998). These differences are due to the formulae for the four base frequencies, with current versions of NHML using the formulae given in the Appendix of Galtier and Gouy's original (1998) paper, while the method used by Galtier and Mouchiroud (1998) used modified equations which account for AT skewness $(A - T)/(A + T)$ and GC skewness $(G - C)/(G + C)$. The original method, which we have used here, is preferable (N. Galtier, personal communication).

Can Mutation Bias Explain Compositional Variation?

The hypothesis to be tested is that compositional variation is due to mutation bias alone. We define two types of polymorphisms: those at which we observe an A or T mutation segregating at a site which was ancestrally G or C, henceforth referred to as a GC \rightarrow AT mutation, and those at which we observe a G or C mutation at a site which was ancestrally A or T, henceforth referred to as an AT \rightarrow GC mutation. If the base composition is stationary and mutation bias solely responsible for base composition bias, then it can be shown that we expect equal numbers of GC \rightarrow AT and AT \rightarrow GC polymorphisms (Eyre-Walker 1997). However, this prediction concerning the numbers of polymorphisms is sensitive to changes in the mutation bias over a timescale of roughly $1/u$ generations, where u is the mutation rate per base pair per generation, since this represents the time required for the composition

to adjust to a change in mutation bias (see Eyre-Walker 1997). Taking u in the mouse as between 10^{-9} (Keightley and Eyre-Walker 2000) and 10^{-8} (Drake et al. 1998) and assuming two mouse generations per year, we have a time scale of between 50 and 500 million years over which the mutation pattern is required to have remained constant. Given that the murid shift must have occurred in the last 100 million years since the common ancestor of mammals (Galtier and Mouchiroud 1998), we need to consider ways of accounting for changing base composition.

The first way to investigate whether mutation bias is responsible for base composition when the composition of the sequences is nonstationary is to determine the equilibrium base composition that the pattern of mutation implies. If the GC content of the sequence is x , then under the mutation bias hypothesis the GC content will move to an equilibrium GC content of

$$\text{predicted equilibrium GC} = xN_{\text{AT} \rightarrow \text{GC}} / ((1 - x)N_{\text{GC} \rightarrow \text{AT}} + xN_{\text{AT} \rightarrow \text{GC}})$$

where the observed numbers of polymorphisms are given by N (see Eq. 9 of Eyre-Walker 1997).

The second way to deal with changing base composition is to perform an alternative test based on the frequency distribution of polymorphisms. If both GC \rightarrow AT and AT \rightarrow GC mutations are neutral, they are expected to be found at the same mean frequency in the population. Thus if we define the mean frequency of GC \rightarrow AT and AT \rightarrow GC polymorphisms to be $F_{\text{GC} \rightarrow \text{AT}}$ and $F_{\text{AT} \rightarrow \text{GC}}$, respectively, then we predict that $F_{\text{GC} \rightarrow \text{AT}} = F_{\text{AT} \rightarrow \text{GC}}$.

Although this frequency test is not completely unaffected by changes in the mutation pattern, the mutation pattern is required to have remained constant for a much shorter time than when the numbers of polymorphisms are considered. A neutral polymorphism is expected to persist for about $4N_e$ generations (Kimura 1983). In the mouse it has been estimated that N_e is around 200,000 (Keightley and Eyre-Walker, personal communication), which means that the polymorphism frequency test is biased only by changes in the mutation bias over the last 800,000 years.

We used polymorphism data from a study which identified SNPs on the basis of comparisons between seven laboratory strains of *Mus musculus domesticus* and one strain of the distinct subspecies *Mus musculus castaneus* (Lindblad-Toh et al. 2000). We used SNPs determined using random genomic STSs, considering only those SNPs for which there was polymorphism within the *M. m. domesticus* strains (data available at <http://www.genome.wi.mit.edu/SNP/mouse>).

The classification of SNPs was carried out on the basis of outgroup comparison by parsimony. For example, if five of the *domesticus* strains are T, and the *castaneus* strain is C, then the SNP was classified as GC \rightarrow AT and the polymorphism frequency was recorded as 2/7. $F_{\text{GC} \rightarrow \text{AT}}$ and $F_{\text{AT} \rightarrow \text{GC}}$ were calculated as the mean frequencies of the two types of polymorphism. We ignored those sites at which parsimony was uninformative. We also recorded the local GC composition using the STS sequences.

Results

Is the Murid Shift Ongoing?

We have inferred the evolution of base composition at the third codon position of protein-coding genes using a data set which allows the comparison of the lineage from the murid ancestor to the mouse-rat ancestor (A in Fig. 1) with the lineages from the mouse-rat ancestor to the present-day mouse and rat sequences (B in Fig. 1). To determine whether the murid shift is ongoing we checked

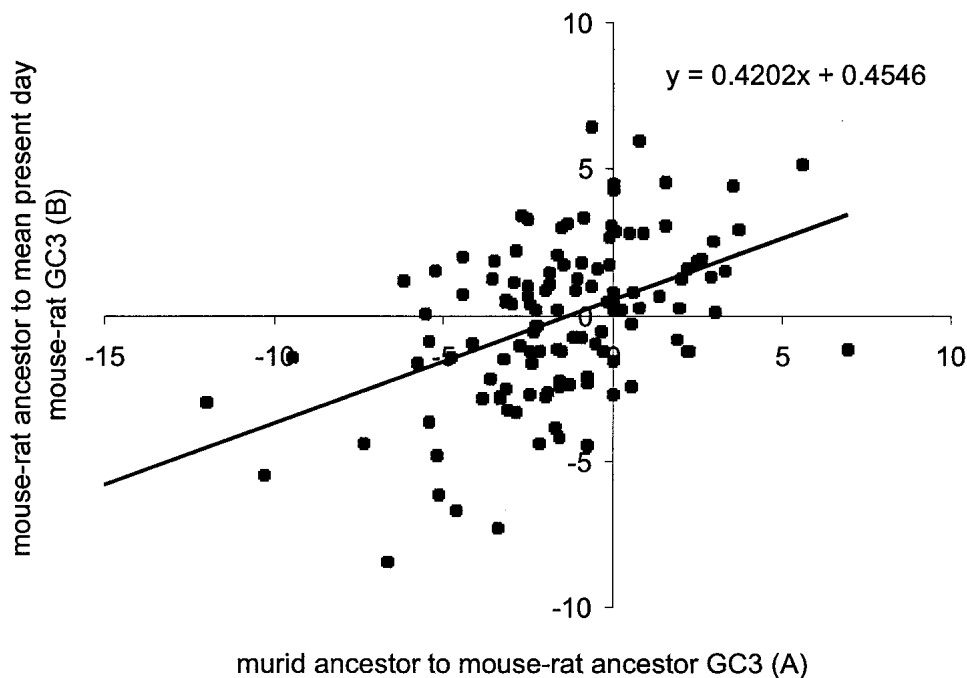


Fig. 2. Ongoing compositional change: comparison of the inferred change in GC3 from the murid ancestor to the mouse-rat ancestor versus the inferred change in GC3 from the mouse-rat ancestor to the present-day mouse and rat sequences for 123 murid genes.

whether there is a correlation between the changes down A and those down B with regard to the G+C content at the third codon position (GC3).

There is indeed a strong positive correlation between the GC3 changes down lineage A and those down lineage B (Spearman's rank correlation, $r^2 = 21\%$, $p < 0.001$; see Fig. 2). In other words, those genes which have decreased in GC3 down lineage A tend to decrease in GC3 down lineage B, and those genes which have increased in GC3 down lineage A tend to increase in GC3 down lineage B. Such a finding is not expected if the base composition is at equilibrium. This result indicates that the pattern of compositional change in lineage A (defined in terms of gene-specific increases and decreases in GC3) is significantly similar to the pattern of compositional change in lineage B. Our result does not unequivocally demonstrate that compositional change is ongoing at present because our GC3 data are insensitive to very recent changes in the evolution of base composition. However, given that the murid shift has taken place in the last 100 million years and that lineage B has evolved over the last 12–14 million years (Galtier and Mouchiroud 1998), the murid shift has definitely been ongoing until the relatively recent evolutionary past.

We also examined the pattern of compositional change. We plotted the change in GC3 from the ancestral murid to the mean of the present-day murids (C in Fig. 1) against the mean present-day murid GC3. Figure 3 shows a significant negative correlation between murid GC3 and the change down lineage C (Spearman's rank correlation, $r^2 = 26\%$, $p < 0.001$); a negative correlation is not expected if the base composition is at equilibrium.

This result confirms previous reports (Galtier and Gouy 1998) of compositional homogenization, with low GC3 sequences tending to increase in GC3 while high GC3 sequences tend to decrease in GC3. The crossover between increasing and decreasing GC3, as inferred by linear regression, is at GC3 = 50% (see Fig. 3).

Can Mutation Bias Explain Compositional Variation?

Given that mutation bias cannot explain the synonymous codon usage of nonrodent mammals (Eyre-Walker 1999; Smith and Eyre-Walker 2001), one possible explanation for the murid shift is that processes of either selection or biased gene conversion, which were present in the ancestral mammal and which persist in nonrodent mammals, were switched off at some point in the evolution of murids. So we have tested whether mutation bias can explain compositional variation in the mouse genome.

Inferring polymorphism on the basis of comparisons between inbred strains of *Mus musculus domesticus* (Lindblad-Toh et al. 2000), we have obtained data for 568 SNPs found in random genomic STSs (see Materials and Methods for details). We have divided the SNPs into three classes based on the local GC content (see Table 1). For each GC class we have calculated the numbers of polymorphisms according to the direction of mutation ($N_{GC \rightarrow AT}$ and $N_{AT \rightarrow GC}$) and the mean frequencies of polymorphisms ($F_{GC \rightarrow AT}$ and $F_{AT \rightarrow GC}$, which can vary between 1/7 and 6/7; see Materials and Methods).

A prediction of the mutation bias hypothesis when the base composition is at equilibrium is that $N_{GC \rightarrow AT} =$

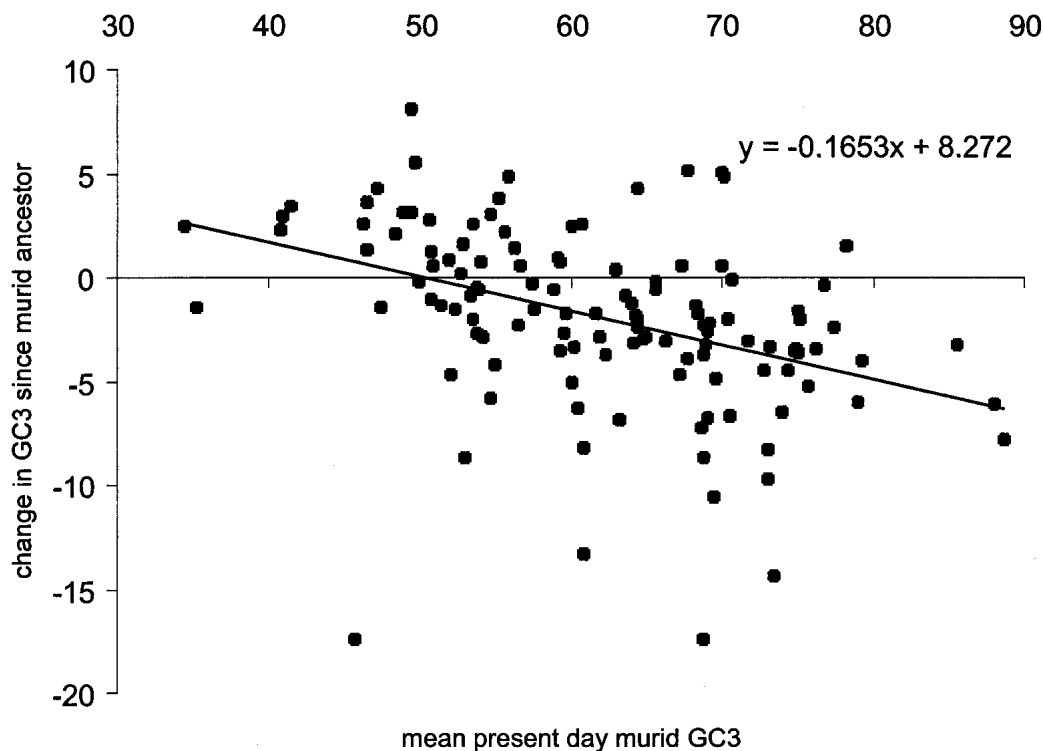


Fig. 3. Compositional homogenization: comparison of the change in GC3 during murid evolution versus the present-day murid GC3.

Table 1. Results of the analyses of nongenic polymorphism in the mouse genome

GC class	Mean GC	No. of SNPs	$N_{GC \rightarrow AT}$	$N_{AT \rightarrow GC}$	$P(N)$	Predicted equilibrium GC	$F_{GC \rightarrow AT}$	$F_{AT \rightarrow GC}$	$p(F)$
All data	43%	568	353	215	$<10^{-8}$	31%	47.3%	50.7%	0.155
<40%	35%	183	103	80	0.075	29%	48.9%	50.0%	0.743
40–50%	45%	286	183	103	$<10^{-5}$	32%	48.6%	51.9%	0.448
>50%	53%	99	67	32	<0.001	35%	40.1%	48.7%	0.167

$N_{AT \rightarrow GC}$ (see Materials and Methods). We find that there are highly significant biases in terms of numbers of polymorphisms, with $N_{GC \rightarrow AT} > N_{AT \rightarrow GC}$ over the full range of GC values [see Table 1; numbers of polymorphism compared with a two-tail binomial test assuming that $N_{GC \rightarrow AT} = N_{AT \rightarrow GC}$ to give $p(N)$]. However, these biases in the numbers of polymorphisms could have been generated by changes in the pattern of mutation bias. We can ask what equilibrium GC would be reached if changes in mutation bias do explain the data (see Materials and Methods). Taking all the data summed across GC classes, we calculate that the mutation bias required to explain this result would cause the GC content to reach equilibrium at 31% (see Table 1). When the data are partitioned into GC classes similar equilibrium GC values are obtained irrespective of GC class (29 to 35%), which suggests that either the entire mouse genome is moving toward very low GC or that selection or biased gene conversion is still acting in the mouse genome.

To examine the issue of selection and biased gene conversion further, we can use polymorphism frequency data. The prediction that $F_{GC \rightarrow AT} = F_{AT \rightarrow GC}$ under neutrality is relatively insensitive to changing composition (see Materials and Methods). However, if selection or biased gene conversion favors either AT or GC alleles, then the neutral prediction does not hold: favored alleles will be found at higher frequencies and disfavored alleles will be found at lower frequencies. The frequency test shows nonsignificant biases from the mutational expectation [see Table 1; polymorphism frequencies compared with a two-tail Mann–Whitney U test to give $p(F)$]. We note that the trend in the frequency data is suggestive of directional selection or biased gene conversion favoring GC alleles and that the frequency test may not be powerful due to the small range of possible polymorphism frequencies. Thus our investigations into the question of whether mutation bias alone can explain compositional variation are not fully conclusive.

Discussion

We have investigated the murid shift by asking two questions. First, is the murid shift ongoing? Our answer to this first question is yes, our analyses of protein-coding genes indicate that the murid shift is continuing. Second, is there any evidence of selection or biased gene conversion affecting base composition in the present-day murids? Our answer to this second question is probably not, since the nongenic polymorphism data are consistent with the mutation bias hypothesis, indicating no significant evidence for either selection or biased gene conversion affecting base composition in the present-day mouse genome.

So what can we conclude about the causes of the murid shift? Unfortunately, inferences concerning the murid shift are complicated by a discrepancy in our results. The polymorphism data, assuming no selection or biased gene conversion, indicate that the GC content is decreasing throughout the mouse genome, but the divergence data indicate that the pattern of compositional change since the murid ancestor has been one of homogenization. The divergence data show that the GC-rich genes have been getting less GC rich but that the GC-poor genes have been getting more GC rich: thus it is the evolution of the GC-poor regions which differs between the divergence and the polymorphism data. We can rule out the explanation that this discrepancy is due to differences in the evolution of genic and nongenic regions, since EST SNPs from the mouse provide evidence of a genomewide decrease in GC similar to that obtained from nongenic SNPs (using SNPs from ESTs with a local GC of less than 40%, we found 15 GC → AT polymorphisms and 8 AT → GC polymorphisms).

We suggest that the explanation for the polymorphism divergence discrepancy may be found in the different time scales over which the polymorphism and divergence data have been generated, roughly 800,000 years for the murid polymorphism (see Materials and Methods) and 12–14 million years for the mouse–rat divergence. It is the originally reported murid shift, causing GC homogenization and occurring up to 100 million years ago, which is revealed by our divergence data. Our polymorphism data can be explained by a second, much more recent, murid shift causing a genomic decrease in GC. Our results may thus be viewed as a preliminary

indication that shifts in genome composition may be more common than previously thought.

Acknowledgments. Thanks go to Nicolas Galtier for advice on the use of the NHML program and to Laurent Duret for discussions. N.G.C.S. was funded by the BBSRC and A.E.-W. is funded by the Royal Society.

References

- Bernardi G (2000) Isochores and the evolutionary genomics of vertebrates. *Gene* 241:3
- Clay O, Caccio S, Zoubak S, Mouchiroud D, Bernardi G (1996) Human coding and noncoding DNA—Compositional correlations. *Mol Phylogenet Evol* 5:2–12
- Drake JW, Charlesworth B, Charlesworth D, Crow JF (1998) Rates of spontaneous mutation. *Genetics* 148:1667–1686
- Duret L, Mouchiroud D, Gouy M (1994) Hovergen—A database of homologous vertebrate genes. *Nucleic Acids Res* 22:2360–2365
- Eyre-Walker A (1997) Differentiating selection and mutation bias. *Genetics* 147:1983–1987
- Eyre-Walker A (1999) Evidence of selection on silent site base composition in mammals: Potential implications for the evolution of isochores and junk DNA. *Genetics* 152:675–683
- Galtier N, Gouy M (1998) Inferring pattern and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol Biol Evol* 15: 871–879
- Galtier N, Mouchiroud D (1998) Isochore evolution in mammals: A human-like ancestral structure. *Genetics* 150:1577–1584
- Keightley PD, Eyre-Walker A (2000) Deleterious mutations and the evolution of sex. *Science* 290:331–333
- Kimura M (1983) *The neutral theory of evolution*. Cambridge University Press, Cambridge
- Lindblad-Toh K, Winchester E, Daly MJ, Wang DG, Hirschhorn JN, Laviolette JP, Ardlie K, Reich DE, Robinson E, Sklar P, Shah N, Thomas D, Fan JB, Gingeras T, Warrington J, Patil N, Hudson TJ, Lander ES (2000) Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature Genet* 24:381–386
- Mouchiroud D, Gautier C, Bernardi G (1988) The compositional distribution of coding sequences and DNA molecules in humans and murids. *J Mol Evol* 27:311–320
- Robinson M, Gautier C, Mouchiroud D (1997) Evolution of isochores in rodents. *Mol Biol Evol* 14:823–828
- Smith NGC, Eyre-Walker A (2001) Synonymous codon bias is not caused by mutation bias in G+C-rich genes in humans. *Mol Biol Evol* 18:892–896
- Thompson JD, Higgins DG, Gibson TJ (1994) ClustalW—Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680
- Wolfe KH, Sharp PM, Li WH (1989) Mutation rates differ among regions of the mammalian genome. *Nature* 337:283–285