Homoplasy for size at microsatellite loci in humans and chimpanzees.

J C Garza and N B Freimer

Genome Res. 1996 6: 211-217
Access the most recent version at doi:10.1101/gr.6.3.211

RESEARCH

Homoplasy for Size at Microsatellite Loci in Humans and Chimpanzees

John Carlos Garza[1,2] and Nelson B. Freimer[2,3]

[1]Department of Integrative Biology, University of California, Berkeley, California 94720; [2]Neurogenetics Laboratory, Department of Psychiatry, University of California, San Francisco, California 94143

Homoplasy (convergence in the size of different alleles) at microsatellite loci was examined by sequencing multiple alleles of two compound microsatellites and single copies of alleles of the same size at two other compound loci in both chimpanzees and humans. At one of the two loci for which multiple alleles were sequenced, extensive homoplasy for size was uncovered both within and between species. At the three loci for which alleles of the same size were examined in the two species, sequencing demonstrated different internal structures. These results confirm theoretical predictions that a certain fraction of mutations at microsatellite loci should produce alleles that are identical in size but differ by a number of mutations. The sequence data reveal a previously unrecognized class of variation at microsatellites and open up the possibility that DNA sequencing may allow the extraction of more information from these loci, thus increasing their power as variable markers for genetic mapping studies. Conversely, the data also indicate that the assumption that alleles of the same size are identical in sequence, which is implicit in several methods of analysis, is violated in some cases. Therefore, caution should be used when employing microsatellites in phylogenetic and other studies in which the individuals being examined are separated by a great number of generations from a common ancestor.

Microsatellite loci are tandemly repeated stretches of a short DNA motif [e.g., $(CA)_n$] which are abundant, extremely polymorphic in length, and relatively evenly spaced in all complex eukaryotic genomes examined to date (Tautz and Schlotterer 1994). They are the primary markers used in most genetic mapping efforts and are also used extensively by population biologists for studies of population structure, parentage, gene flow, and hybridization (Bruford and Wayne 1993). Microsatellite loci are often conserved over long periods of evolutionary time, which has led to their employment for phylogenetic reconstruction using distance measures calculated from pair–wise differences and allele frequencies (Bowcock et al. 1994; Goldstein et al. 1995). Many of these uses rest on the assumption that alleles of the same size are identical in sequence and that they differ by no mutations. The polymorphism at microsatellite loci is assayed most commonly by simply examining length variation of an amplified PCR product containing the repeat region on a high-resolution acrylamide or agarose gel. Size differences among alleles are assumed to be attributable to differences in the integral number of repeats.

Previous studies have shown that the microsatellite mutation process, at least for dinucleotide loci, is described adequately by the one-step mutation model (Ohta and Kimura 1973) or one of its generalizations (Shriver et al. 1993; Valdes et al. 1993; Di Rienzo et al. 1994), in which all or most mutations change allele size by a single repeat unit. The bidirectional nature of microsatellite mutation, as both described by these models and directly observed (Weber and Wong 1993), indicate that convergence in size should occur at these loci. Additionally, it has been found previously that differences in average repeat size between chimpanzees and humans tend to be smaller than would be expected if evolution at these loci was neutral and unconstrained (Garza et al. 1995). These observations, coupled with high rates of mutation at microsatellite loci, raise the possibility that homoplasy, an external similarity masking significant evolutionary differences, may exist for size at microsatellite loci in many populations. That is, copies of a locus that share the same size may differ by an unknown number of mutations. However, in most cases it would be impossible to detect homoplasy di-

[3]Corresponding author.
E-MAIL nbfr@itsa.uscf.edu; FAX (415)476-7389.

rectly at simple microsatellite loci (Fig. 1A). For example, if copies A and B of a locus are identical by descent and thus are initially the same size, and copy A undergoes two gains of one repeat unit while copy B undergoes a loss of one repeat unit and then a gain of three, they would again share the same size. These alleles would thus appear identical by most survey methods, including DNA sequencing, although they are separated by four mutations.

Compound microsatellite loci, however, offer an opportunity to detect convergence in size. Compound microsatellites contain at least two repeated regions found directly adjacent to one another, and differences between alleles may be attributable to variation in both of the repeat regions (Garza et al. 1995). Mutations in the two distinct repeat regions could create alleles of the same size that have different combinations of repeats numbers and are not identical in sequence (Fig. 1B). We have sequenced multiple copies of two compound microsatellite loci and single copies of two additional compound loci in both the human and chimpanzee. Our results show that homoplasy for size occurs at these loci and that this homoplasy masks a class of variation at microsatellite loci that previously has been overlooked for the most part. DNA sequencing may allow the extraction of additional information from these already important genetic markers by permitting definition of alleles on the basis of sequence rather than size. These results should spur the use of sequence-level variation at microsatellite loci by those who aim to further refine the associations caused by genetic linkage. They should suggest also a note of caution for studies that rely entirely on size variation of microsatellite loci to infer identity by descent of alleles.

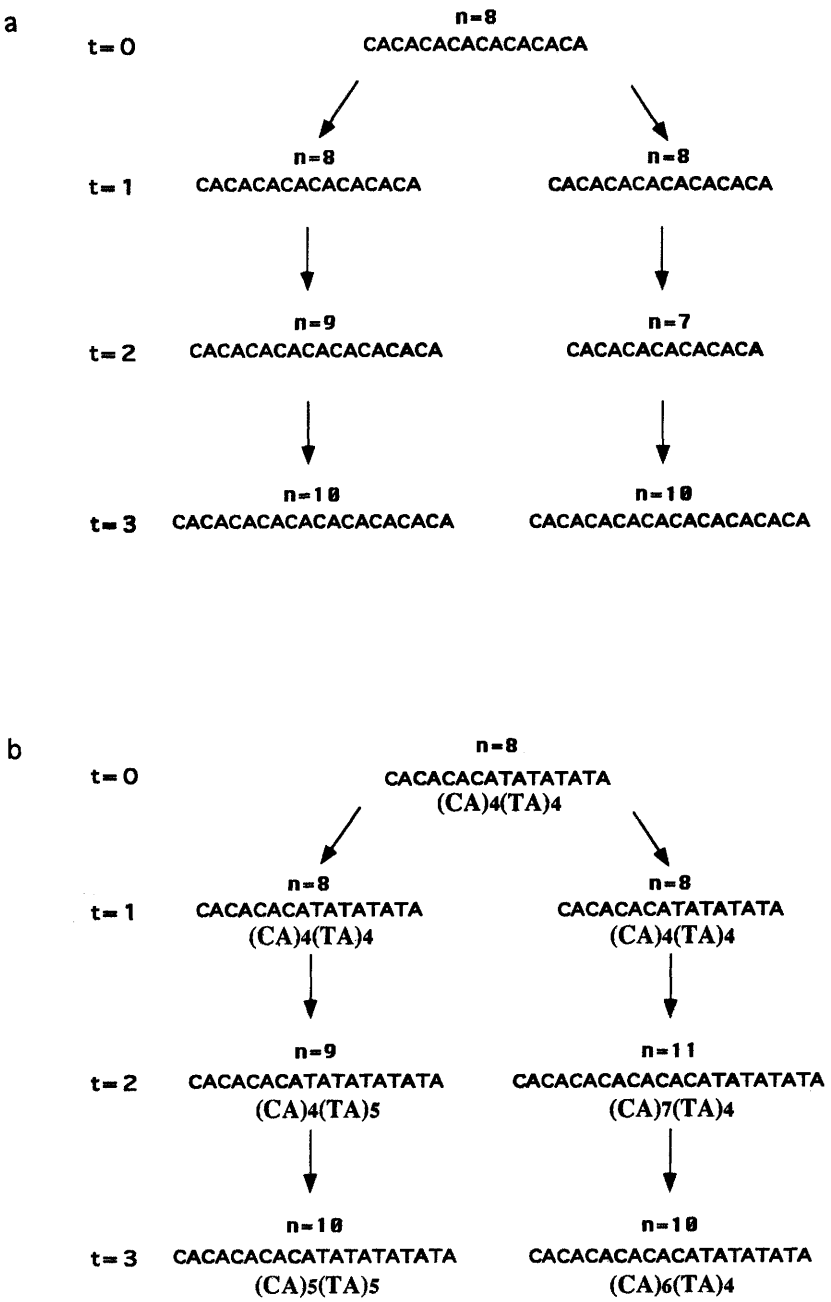

Figure 1 Diagram of convergence in size under the two-phase model of mutation (Di Rienzo et al. 1994). (*a*) Convergence caused by mutations in a single repeat region. (*b*) Convergence caused by mutations in two separate repeat regions.

RESULTS

The sequences of 14 chimpanzee alleles and 9 human alleles at locus Mfd 59 are shown in Table 1. The allelic (size) class frequency distributions of this locus in the two species are shown in Figure 2. The major difference between al-

Table 1. Repeat Region Sequence of Copies of Locus Mfd 59 (D4S174) in the Chimpanzee and Human

Chimpanzee		Human	
size (bp)	sequence	size (bp)	sequence
180	$(CA)_1(TA)_{20}(CA)_{16}$	194	$(CA)_5(TN)_{11}(CA)_{28}$
172	$(CA)_1(TA)_{20}(CA)_{12}$ *	184	$(CA)_5(TN)_{11}(CA)_{23}$
172	$(CA)_1(TA)_{21}(CA)_{11}$ *	182	$(CA)_5(TN)_{12}(CA)_{21}$ ^
170	$(CA)_1(TA)_{19}(CA)_{12}$ #	182	$(CA)_5(TN)_{14}(CA)_{19}$ ^
170	$(CA)_1(TA)_{18}(CA)_{13}$ #	180	$(CA)_5(TN)_{13}(CA)_{19}$ °
168	$(CA)_1(TA)_{18}(CA)_{12}$	180	$(CA)_5(TN)_{13}(CA)_{19}$ °
164	$(CA)_1(TA)_{12}(CA)_{16}$	180	$(CA)_5(TN)_{13}(CA)_{19}$ °
158	$(CA)_1(TA)_{11}(CA)_{14}$ †	180	$(CA)_5(TN)_{14}(CA)_{18}$ °
158	$(CA)_2(TA)_{10}(CA)_{14}$ †	180	$(CA)_5(TN)_{14}(CA)_{18}$ °
156	$(CA)_1(TA)_{11}(CA)_{13}$		
156	$(CA)_1(TA)_{11}(CA)_{13}$		
156	$(CA)_1(TA)_{11}(CA)_{13}$		
154	$(CA)_1(TA)_{10}(CA)_{13}$		
154	$(CA)_1(TA)_{10}(CA)_{13}$		

(*) (#) (†) (^) (°) Alleles of the same size with different internal structures.

leles at Mfd 59 in chimpanzees and humans is that the perfect AT repeat of the chimpanzee locus is an AT-rich region in humans that differs from perfection by three transversions (Garza et al. 1995). Another difference at this locus between the two species is that the region containing five CA repeats in the human is reduced to one or two CA repeats in the chimpanzee. Three of the five allelic classes in the chimpanzee for which multiple sequences were determined had examples of homoplasy in size. The two allelic classes in humans for which multiple sequences were determined also had examples of homoplasy. In contrast, Mfd 75, for which sequences are shown in Table 2, had no examples of homoplasy in size for the one allelic class for which multiple sequences were determined. One apparently fixed-base substitution in the flanking sequence was found between chimps and humans at each of the two loci (data not shown). The extent of homoplasy found at these loci is summarized in Table 3. Overall, homoplasy for size was found in more than half (5/8) of the allelic classes for which more than one sequence was determined. Single sequences of alleles of the same size at two additional compound loci were both found to have different internal structures in the human and chimpanzee (Table 4). The alleles of Mfd 38 differ by at least two mutations, whereas the differences between alleles at Mfd 104 are more complex. The expansion of the 5′ CA region and the two transitions (A → G and C → T) are shared by other chimpanzee alleles sequenced (data not shown).

DISCUSSION

We have shown that convergence in size occurs at microsatellite loci and that it may be a relatively common phenomenon in these increasingly widely used genetic markers. We have demonstrated the existence of homoplasy for size both within and between species through sequencing of compound microsatellite alleles. At one microsatellite locus (Mfd 59) homoplasy was present at five of eight allelic classes examined in the two species, thereby substantially increasing the number of alleles recognized at this locus. At the three loci for which alleles of the same size were examined in the two species, all had different sequence-level structures. Our results thus indicate that DNA sequencing of compound microsatellites can permit differentiation of alleles that although identical in size, are not identical in sequence. These loci represent a substantial proportion of all microsatellites [11% in one collection of dinucleotide repeats (Weber 1990)] and thus represent an untapped reservoir of potentially useful genetic variation. By designating alleles based on sequence and not on size, this previously unrecognized variation may increase the power of microsatellites in genetic mapping efforts. Conversely, such variation may pose problems for the use of microsatellites in some types of analysis, as relationships between alleles will be obscured as the number of generations separating them increases.

Such homoplasy for size is a prediction of the one-step mutation model and its generalizations. The one-step mutation process working independently in two different repeat regions of a compound locus could cause convergence at two lev-

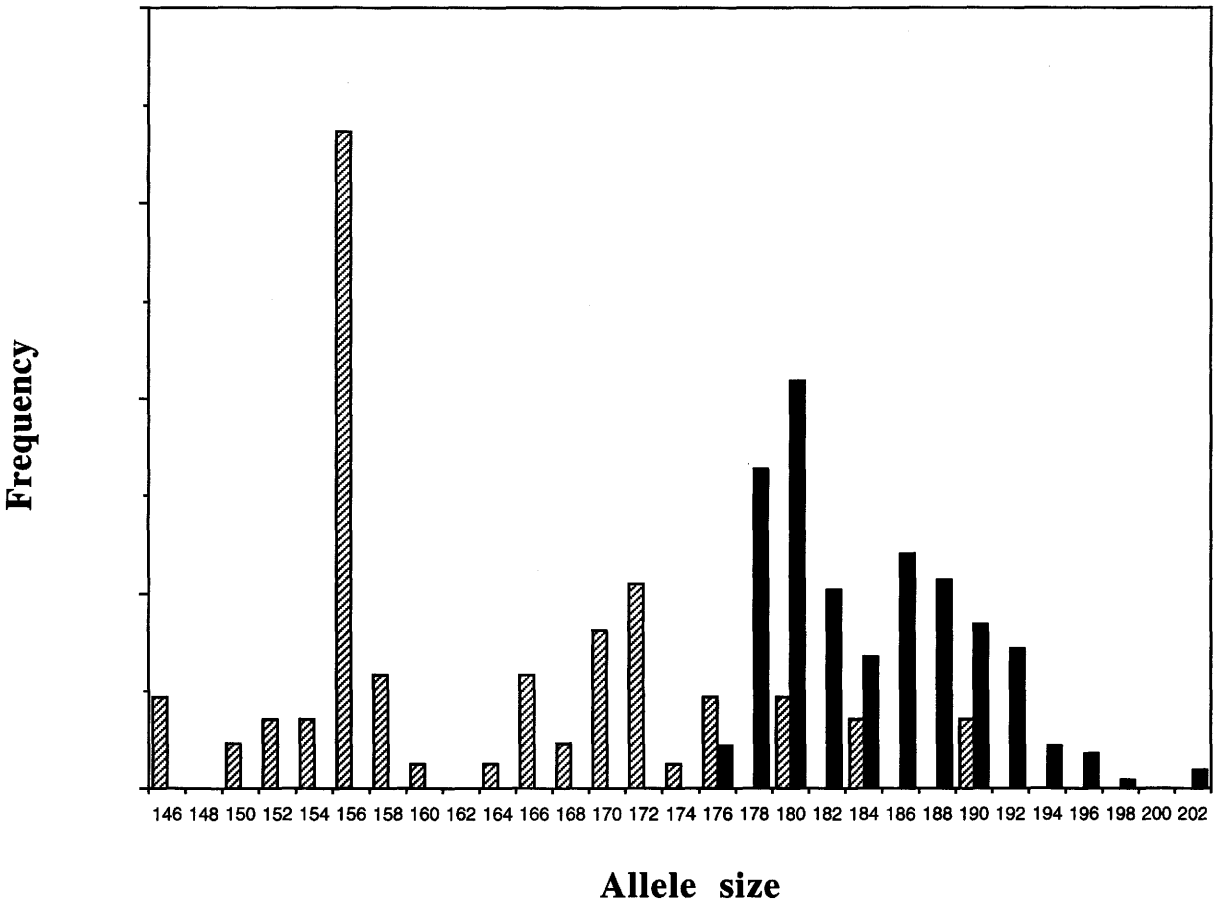

Figure 2 Allelic class frequency distributions of Mfd 59 (D4S174) in the chimpanzee (hatched bars) and human (solid bars). Allele size is in base pairs.

els. The first level, convergence within a repeat region (Fig. 1A), would not be directly detectable. The second level, however, convergence caused by different combinations of repeats in the two repeat regions (Fig. 1B), is directly observable through DNA sequencing.

Although sequencing will not uncover homoplasy for size at simple microsatellite loci, it might still be possible to detect homoplasy by examining patterns of disequilibrium with closely linked polymorphisms, such as base substitutions or insertion/deletion elements (indels). M.-C. Grimaldi, P. Avoustin, and B. Crouau-Roy (unpubl.) found that a 6-bp indel in the region directly adjacent to a CA repeat region in the human leukocyte antigen region resulted in alleles of the same size that differed in the presence or absence of the element. Estoup et al. (1995) have shown recently that extensive homoplasy for size is present among different subspecies of honeybee at an imperfect microsatellite locus by examining the position of imperfections within the repeat region. Variation in the length of the sequence between the repeat region and the PCR primer may also be an important source of homoplasy at microsatellite loci.

In theory, it would be possible to estimate the amount of homoplasy in a sample of copies of a microsatellite locus. It would, however, require complete information about the mutation process of the locus in question. This would have to include rate, the divergence time of the copies of the locus in the sample, the selective history of the locus, the percentage of mutations that are not single step, and any potential directionality of mutation. In practice, it is difficult to obtain adequate information about all of these parameters. For cases in which identity of alleles is crucial, the most appropriate strategy might be to estimate empirically the level of homoplasy by

Table 2. Repeat Region Sequence of Copies of Locus Mfd 75 (D12S59) in the Chimpanzee and Human

Chimpanzee		Human	
size (bp)	sequence	size (bp)	sequence
150	$(AC)_{10}(TC)_{10}(TG)_1(TC)_4$	180	$(AC)_{16}(TC)_{19}(TG)_1(TC)_4$
148	$(AC)_{10}(TC)_9(TG)_1(TC)_4$	174	$(AC)_{20}(TC)_{12}(TG)_1(TC)_4$
148	$(AC)_{10}(TC)_9(TG)_1(TC)_4$	172	$(AC)_{19}(TC)_{12}(TG)_1(TC)_4$
148	$(AC)_{10}(TC)_9(TG)_1(TC)_4$		
148	$(AC)_{10}(TC)_9(TG)_1(TC)_4$		
146	$(AC)_9(TC)_9(TG)_1(TC)_4$		
144	$(AC)_{10}(TC)_7(TG)_1(TC)_4$		

DNA sequencing. It should be stressed, however, that DNA sequencing will still underestimate the level of homoplasy at a compound locus. Alleles could still have the same combination of repeat numbers in the different repeat regions (and therefore be identical in sequence) but not be identical by descent.

The fact that no homoplasy was uncovered at Mfd 75 may not be attributable to the fact that no homoplasy is present at this locus but only that none was uncovered in our small sample of sequences. Alternatively, the absence of homoplasy may be attributable to a relatively recent common ancestor for all of the copies within the allelic class examined or to some recent selective event affecting this locus (Slatkin 1995). Samples were not available for the individuals comprising the two groups that shared alleles of the same size and thus it was impossible to examine whether interspecific homoplasy is present at Mfd 75.

The existence of two disjunct classes of repeat numbers (10–12 and 18–21) for the TA repeat region in the copies of Mfd 59 from the chimpanzee is consistent with the two-phase mutation model introduced by Di Rienzo et al. (1994). This suggests that the two modes of the chimpanzee allele frequency distribution at Mfd 59 (Fig. 1) are the result of a mutation of large effect from an allele in one of the size classes followed by a number of single-step mutations. Sequence data from closely related species should allow determination of whether or not this is the case. It is interesting also to note that in some cases, alleles that are separated in size by 2 bp differ by more than one mutation.

Although homoplasy for size is probably a relatively widespread phenomenon, compound loci should on average have higher levels of homoplasy than simple loci because of mutations accumulating in more than one repeat region. Thus, if the aim is to extract the maximum amount of polymorphic information from particular marker loci, compound loci are perhaps superior to simple loci, as some of this additional variation may be detectable by sequencing. For example, in a linkage mapping study a marker that appears to be uninformative when alleles are defined on the basis of size may become informative when sequence-level variation is considered. Association and linkage disequilibrium mapping studies (Lander and Botstein 1986; Houwen et al. 1994) also may benefit by allowing the detection of previously unrecognized disequilibrium.

However, the existence of homoplasy for size at microsatellite loci also poses problems for a number of different types of studies in which these loci are employed. Although the probability of convergence occurring on the time scale of pedigree-based linkage mapping is very small, it may be an important source of error in any study in which a population-based approach is used, such as phylogeny estimation, when only allele size is considered. Studies in which microsatellites have been used for phylogeny reconstruction have employed distance measures based on allele frequencies and pair–wise differences (Bowcock et al. 1994; Goldstein et al. 1995). Some microsatellite loci are ancient in origin, and the fact

Table 3. Extent of Homoplasy for Loci Mfd59 and Mfd75

	Species	
Locus	chimpanzee	human
Mfd 59	3/5	2/2
Mfd 75	0/1	N.A.

Extent of homoplasy found at the two loci for which multiple copies were examined. The ratios are the number of allelic classes for which homoplasy was found out of all allelic classes examined. (N.A.) Not applicable.

Table 4. Repeat Region Sequence of Alleles of the Same Size from the Microsatellite Loci Mfd 38 (D4S175) and Mfd 104 (D8S164) in Both the Chimpanzee and Human

Mfd 38

Human 124bp	Chimp 124bp
(TA)11(CA)15	(TA)10(CA)16

Mfd 104

Human 169bp	Chimp 169bp
(TA)10(CA)1(A)6(TA)7T(TA)5(CA)14	(TA)13(CA)5(A)3G(TA)6T(TA)5(CA)6TA(CA)2

that large differences in the mean number of repeats and in allele frequency distributions are not found between species that have been evolving independently for millions of years indicates that there are probably some constraints on allele size (Deka et al. 1994; Fitzsimmons et al. 1995; Garza et al. 1995). Such a constraint, coupled with the high mutation rates of microsatellites (estimated between 10^{-2}–10^{-5} locus/generation; Weber and Wong 1993), support the assertion that some mutational events must create alleles that do not differ in size from previously existing alleles and that any information about genealogical relationships will be lost quickly even without substantial divergence of allele frequency distributions. The probability that homoplasy is present at a microsatellite locus will increase, up to a certain point, with increasing coalescence time of the copies in the sample, until all patterns of similarity attributable to hierarchical levels of relationship are lost. The differences between the basic structure of alleles of the same size at Mfd 59 and Mfd 104 in the chimpanzee and human suggest that actually they are separated by many mutations and that a comparison of this locus based on size would provide no real genealogical information at this level of evolutionary divergence. Therefore, caution should be used when employing microsatellites in analyses in which the units under study are separated by significant amounts of evolutionary time. For example, relationships between populations within a species may be resolved with microsatellite polymorphisms, but relationships among species or genera will most likely be obscured.

METHODS

Genomic DNA was extracted from blood collected from unrelated individuals. All of the human DNA samples were from a group of African individuals studied by Di Rienzo et al. (1994), and the chimpanzee DNA samples were from the colony at Yerkes Regional Primate Research Center (Atlanta, GA). The taxonomic affinities of the Yerkes colony animals are not known, and thus, it is possible that individuals examined in this study belong to different subspecies or even different species (Morin et al. 1994). PCR was performed as described previously (Garza et al. 1995) except that the total reaction volume was 50 ml. PCR products were ligated directly into Invitrogen TA cloning vector, or purified and ligated into Promega pGEM-T vector, and then cloned into JM109-competent cells following the manufacturer's specifications. Plasmid DNA was isolated from positive colonies, and the insert sequenced with both the M13 forward and reverse universal primers using the cycle sequencing protocol from Promega. The sequence of human allele 2 (184 bp) at Mfd 59 (D4S174) was extracted from GenBank (accession no. X54584), as were the sequences of human allele 1 (180 bp) at Mfd 75 (D12S59) (accession no. M83618) and the human allele at Mfd 38 (D4S175) (accession no. M84934). The sequences described in the text have been deposited in GenBank (accession nos. U48299–U48332). The allele frequencies in Figure 2 were calculated from 48 chromosomes of unrelated chimpanzees and 225 chromosomes of unrelated human individuals from Africa and Europe.

ACKNOWLEDGMENTS

We thank M. Slatkin, L. Bull, and B. Crouau-Roy for discussion and advice. J.C.G. was supported by the Ford Foundation. This work was supported by grants from the National Institutes of Health to N.B.F. and M. Slatkin. The sequence data described in this paper have been submitted to GenBank under accession numbers U48299–448332.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Bowcock, A.M., A. Ruiz-Linares, J. Tomfohrde, E. Minch, J.R. Kidd, and L.L. Cavalli-Sforza. 1994. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368:** 455–457.

Bruford, M.W. and R.K. Wayne. 1993. Microsatellites and their application to population genetic studies. *Curr. Opin. Genet. Dev.* **3:** 939–943.

Deka, R., M.D. Shriver, L.M. Yu, L. Jin, C.E. Aston, R. Chakraborty, and R.E. Ferrell. 1994. Conservation of human chromosome 13 polymorphic microsatellite (CA)(n) repeats in chimpanzees. *Genomics* **22:** 226–230.

Di Rienzo, A., A.C. Peterson, J.C. Garza, A.M. Valdes, M. Slatkin, and N.B. Freimer. 1994. Mutational processes of simple-sequence repeat loci in human populations. *Proc. Natl. Acad. Sci.* **91:** 3166–3170.

Estoup, A., C. Tailliez, J.M. Cornuet, and M. Solignac. 1995. Size homoplasy and mutational processes of interrupted microsatellites in two bee species, *Apis mellifera* and *Bombus terrestris* (Apidae). *Mol. Biol. Evol.* **12:** 1074–1084.

Fitzsimmons, N.N., C. Moritz, and S.S. Moore. 1995. Conservation and dynamics of microsatellite loci over 300 million years of marine turtle evolution. *Mol. Biol. Evol.* **12:** 432–440.

Garza, J.C., M. Slatkin, and N.B. Freimer. 1995. Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. *Mol. Biol. Evol.* **12:** 594–603.

Goldstein, D.B., A.R. Linares, L.L. Cavalli-Sforza, and M.W. Feldman. 1995. Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc. Natl. Acad. Sci.* **92:** 6723–6727.

Houwen, R.H.J., S. Baharloo, K. Blankenship, P. Raeymakers, J. Juyn, L.A. Sandkuijl, and N.B. Freimer. 1994. Genome screening by searching for shared segments: Mapping a gene for benign recurrent intrahepatic cholestasis. *Nature Genet.* **8:** 380–386.

Lander, E.S. and D. Botstein. 1986. Mapping complex genetic traits in humans: New approaches using a complete RFLP linkage map. *Cold Spring Harbor Symp. Quant. Biol.* **51:** 49–52.

Morin, P.A., J.J. Moore, R. Chakraborty, L. Jin, J. Goodall, and D.S. Woodruff. 1994. Kin selection, social structure, gene flow, and the evolution of chimpanzees. *Science* **265:** 1193–1201.

Ohta, T. and M. Kimura. 1973. The model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a genetic population. *Genet. Res.* **22:** 201–204.

Shriver, M.D., L. Jin, R. Chakraborty, and E. Boerwinkle. 1993. VNTR allele frequency distributions under the stepwise mutation model—A computer simulation approach. *Genetics* **134:** 983–993.

Slatkin, M. 1995. Hitchhiking and associative overdominance at a microsatellite locus. *Mol. Biol. Evol.* **12:** 473–480.

Tautz, D. and C. Schlotterer. 1994. Simple sequences. *Curr. Opin. Genet. Dev.* **4:** 832–837.

Valdes, A.M., M. Slatkin, and N.B. Freimer. 1993. Allele frequencies at microsatellite loci: The stepwise mutation model revisited. *Genetics* **133:** 737–749.

Weber, J.L. 1990. Informativeness of human $(dC\text{-}dA)_n\cdot(dG\text{-}dT)_n$ polymorphisms. *Genomics* **7:** 524–530.

Weber, J.L. and C. Wong. 1993. Mutation of human short tandem repeats. *Hum. Mol. Genet.* **2:** 1123–1128.

Received November 29, 1995; accepted in revised form February 6, 1996.