Low Nucleotide Diversity in Man

Wen-Hsiung Li and Lori A. Sadler

Center for Demographic and Population Genetics, University of Texas Health Science Center, P.O. Box 20334, Houston, Texas 77225

Manuscript received January 14, 1991
Accepted for publication June 6, 1991

ABSTRACT

The nucleotide diversity (π) in humans is studied by using published cDNA and genomic sequences that have been carefully checked for sequencing accuracy. This measure of genetic variability is defined as the number of nucleotide differences per site between two randomly chosen sequences from a population. A total of more than 75,000 base pairs from 49 loci are compared. The DNA regions studied are the 5′ and 3′ untranslated regions and the amino acid coding regions. The coding regions are divided into nondegenerate sites (*i.e.*, sites at which all possible changes are nonsynonymous), twofold degenerate sites (*i.e.*, sites at each of which one of the three possible changes is synonymous) and fourfold degenerate sites (*i.e.*, sites at which all three possible changes are synonymous). The π values estimated are, respectively, 0.03 and 0.04% for the 5′ and 3′ UT regions, and 0.03, 0.06 and 0.11% for nondegenerate, twofold degenerate and fourfold degenerate sites. Since the highest π value is only 0.11%, which is about one order of magnitude lower than those in Drosophila populations, the nucleotide diversity in humans is very low. The low diversity is probably due to a relatively small long-term effective population size rather than any severe bottleneck during human evolution.

HOW much genetic variability exists in human populations? We certainly have already learned a great deal about this question, for it has been known since HARRIS (1966) that human populations, like many other natural populations, contain considerable genetic variability (see, *e.g.*, LEWONTIN 1974; HARRIS, HOPKINSON and EDWARDS 1977). For example, the average heterozygosity for *Homo sapiens* estimated from electrophoretic data (based on 121 loci) is 14%, which is the same as that for *Drosophila pseudoobscura* estimated from 46 loci (see data compiled by NEI and GRAUR 1984). However, can this observation at the protein level be extrapolated to the nucleotide level? Although recombinant DNA techniques have long provided us with the necessary tools for addressing this question, no extensive study seems to have been conducted in any human populations. Fortunately, we have been able to obtain information for this question from published human DNA sequence data. Our analysis indicates that at the nucleotide level the genetic variability in human populations is one order of magnitude lower than that in *D. pseudoobscura*. This raises an intriguing question: Why should the genetic variabilities in the two species differ so greatly at the nucleotide level but not at the protein level?

MATERIALS AND METHODS

The DNA sequence data used in this study were obtained from the literature or GenBank. There are many human genes, particularly their coding regions, that have been sequenced from two or more individuals. However, the majority of these sequences are not suitable for the present purpose because their sequence accuracy has not been examined carefully. To minimize the effect of sequencing errors, we chose genes according to the following criteria. A pair of sequences have been obtained in the same laboratory and have been carefully checked against each other; they were usually one cDNA and one genomic sequence and the cDNA and genomic libraries used were constructed from different individuals. That is, we first investigated the availability of multiple sequences for a locus (gene) and then went back to the original papers to check whether any two of them were from the same laboratory and whether the original authors have carefully compared the sequences to each other. We included only sequences that were accompanied by published comments by the authors confirming that any sequence differences had been checked and corrected (if an error was detected) or checked and confirmed as being different. In the few cases where the sequences originated from different laboratories, nucleotide differences were documented by at least one of the two groups or by other evidence (such as the availability of other independently published sequences). Since all the nucleotide differences between sequences were stated in the original papers, the data we compiled presumably include no or few recording errors. For comparison, we also studied another random set of data for which the sequences had not been checked against each other.

The measure of genetic variability used is the nucleotide diversity, which is defined as the number of differences per nucleotide site between two randomly chosen sequences from a population (NEI and LI 1979). This measure is the same as the proportion of different nucleotides between two random sequences. When more than two sequences are available, the average nucleotide diversity is computed as the arithmetic mean of all pairwise comparisons. We consider coding and noncoding regions separately (see Table

TABLE 1

Nucleotide diversity in humans

Gene[a]	Noncoding regions		Coding regions				Refs.[b]
				2-fold deg.			
	5′UT	3′UT	Nondegenerate	Nonsynonymous	Synonymous	4-fold degenerate	
Acid glycoprotein-α_1	0/78	0/143	0/383	0/134	0/134	0/83	1
Adenosine deaminase	0/72	0/311	0/701	0/212	0/212	2/173	2, 3
α-Amylase (salivary)	0/199	0/30	0/1,017	0/301	0/301	0/215	4, 5
Aldose reductase	NA	0/371	0/612	0/200	0/200	0/133	6, 7
Androgen receptor	0/77	0/139	0/1,773	0/510	0/510	0/471	8, 9
Angiogenin	NA	0/175	0/271	0/100	0/100	0/67	10
Angiotensinogen	0/39	0/602	1/919	0/268	0/268	0/265	11, 12
Apolipoprotein A-I	0/86	0/57	0/508	0/164	0/164	0/126	13
Apolipoprotein A-II	0/9	0/112	0/188	0/60	0/60	0/49	14, 15
Apolipoprotein E	0/61	0/142	1/571	0/158	3/158	0/168	16, 17
Apoferritin H	0/91	0/161	0/349	0/134	0/134	0/66	18, 19
Calcitonin/CGRP	0/74	NA	0/235	0/80	0/80	0/66	20
Cathepsin G	0/8	0/81	0/466	0/167	0/167	0/129	21, 22
Complement C1 inhibitor	NA	0/264	0/625	0/190	0/190	0/154	23
Elongation factor-1α	0/53	0/295	0/911	0/245	0/245	0/227	24
Erythropoietin	0/179	0/565	0/359	0/103	0/103	0/114	25
Factor VIII	0/109	0/1,800	0/4,558	1/1,527	0/1,527	0/965	26, 27
Factor IX	NA	0/1,389	1/900	0/292	0/292	0/188	28, 29
Factor X	NA	NA	0/927	0/300	0/300	0/201	30
Fibrinogen-γ	0/80	0/241	1/858	0/288	1/288	2/162	31, 32
Gdx	0/35	0/1815	0/321	0/113	0/113	0/91	33
Granulocyte colony stimulating factor	0/31	0/853	0/389	0/112	0/112	0/117	34, 35
Hepatic lipase	0/4	0/46	1/964	0/301	0/301	1/229	36, 37
Interleukin-1β	0/87	0.7/599	0/525	0/175	0/175	0/104	38–40
Interleukin-5	0/44	0.7/357	0/245	0/97	0/97	0/57	41–43
Keratin-18	0/47	1/68	0/811	0/268	0/268	0/208	44, 45
α-Lactalbumin	NA	0/272	0/271	0/106	0/106	0/46	46, 47
Lactate dehydrogenase A	0/25	0/565	0/649	0/195	0/195	0/149	48, 49
Lactate dehydrogenase B	0/7	0/200	0/655	0/195	0/195	0/149	50, 51
Ly1-1	0/258	0/286	0/496	0/130	0/130	0/172	52
Pancreatic polypeptide	0/56	0/76	0.5/173	0/53	0/53	0/56	53–56
Parathyroid hormone	0/74	0/348	0/221	0/72	0/72	0/49	57, 58
Phosphoglycerate kinase	0/79	0/434	0/813	0/230	0/230	0/205	59, 60
Phosphoglycerate mutase	0/35	0/36	1/495	0/141	0/141	1/121	61, 62
Plasmin inhibitor-α_2	NA	1/729	1/933	0/277	0/277	0/254	63, 64
Prolactin	0/4	0/145	0/430	0/141	0/141	0/107	65, 66
Protein C	0/98	0/360	0/899	0/274	0/274	0/207	67, 68
Renin	0/42	0/172	0/795	0/217	0/217	0/203	69, 70
Ribosomal protein S14	0/33	0/45	0/294	0/74	0/74	0/85	71
Steroid 21-hydroxylase	0/32	0/492	0/932	0/287	0/287	0/260	72
Superoxide dismutase (Cu/Zn)	NA	0/94	0/297	0/87	0/87	0/75	73, 74
Thrombomodulin	0/150	0/1,117	0/1,115	0/291	0/291	0/316	75, 76
Tissue plasminogen activator	0/218	0/755	0/1,077	0/358	1/358	1/248	77
Thymidine kinase	1/57	0/659	0/450	0/133	0/133	0/116	78, 79
Triosephosphate isomerase	0/367	0/718	0/487	0/129	0/129	1/128	80, 81
Tumor antigen p53	0/135	NA	1/762	0/210	0/210	0/204	82, 83
Tumor necrosis factor	0/152	0/789	0/440	0/130	0/130	0/126	84, 85
Vasoactive intestinal polypeptide	0/176	4/767	0/325	0/108	0/108	0/74	86
Von Willebrand factor	0/163	0/94	2/1,474	0/450	0.5/450	1.3/359	87–91
Total	1/3,624 0.0003 ± 0.0003[c]	7.4/19,769 0.0004 ± 0.0001	10/34,869 0.0003 ± 0.0001	1/10,787 0.0001 ± 0.0001	5.5/10,787 0.0005 ± 0.0002	9.3/8,537 0.0011 ± 0.0004	

[a] The numbers of sequences used are three for each of the genes for acid glycoprotein α1, interleukin-1β, and interleukin 5, two for the pancreatic polypeptide gene, eight for the von Willebrand factor gene, and two for each of the other genes.

[b] 1. DENTE, CILIBERTO and CORTESE (1985); 2. ADRIAN, WIGINTON and HUTTON (1984); 3. WIGINTON *et al.* (1986); 4. NAKAMURA *et al.* (1984); 5. NISHIDE *et al.* (1986); 6. GRAHAM *et al.* (1989); 7. BOHREN *et al.* (1989); 8. LUBAHN *et al.* (1988); 9. LUBAHN *et al.* (1989); 10. KURACHI *et al.* (1985); 11. KAGEYAMA, OHKUBO and NAKANISHI (1984); 12. KUNAPULI and KUMAR (1986); 13. SEILHAMER *et al.* (1984); 14. MOORE *et al.* (1984); 15. TSAO *et al.* (1985); 16. MCLEAN *et al.* (1984); 17. PAIK *et al.* (1985); 18. COSTANZO *et al.* (1984); 19. COSTANZO *et al.* (1986); 20. JONAS *et al.* (1985); 21. SALVESEN *et al.* (1987); 22. HOHN *et al.* (1989); 23. CARTER, DUNBAR and FOTHERGILL (1988); 24. UETSUKI *et al.* (1989); 25. JACOBS *et al.* (1985); 26. WOOD *et al.* (1984); 27. GITSCHIER *et al.* (1984); 28. MCGRAW *et al.* (1985); 29. YOSHITAKE *et al.* (1985); 30. LEYTUS *et al.* (1986); 31. CHUNG and DAVIE (1984); 32. RIXON, CHUNG and DAVIE (1985); 33. TONIOLO, PERSICO and ALCALAY (1988); 34. NAGATA *et al.* (1986a); 35. NAGATA *et al.* (1986b); 36. DATTA *et al.* (1988); 37. CAI *et al.* (1989); 38. CLARK *et al.* (1986); 39. BENSI *et al.* (1987); 40. NISHIDA *et al.* (1987); 41. AZUMA *et al.* (1986); 42. TANABE *et al.* (1987); 43. Campbell *et al.* (1987); 44.

1). The noncoding regions are divided into the 5′ and 3′ untranslated (UT) regions; flanking (untranscribed) regions are not considered because of the paucity of data. In coding regions, a site is labeled nondegenerate if all possible changes at that site are nonsynonymous (amino acid-changing), twofold degenerate if one of the three possible changes is synonymous, and fourfold degenerate if all three possible changes are synonymous. The calculation can easily be done by using the computer program of LI, WU and LUO (1985).

RESULTS

Nucleotide diversity: We have been able to find 49 genes that are suitable for the present purpose. Table 1 shows the results of our analysis. In most of the cases no difference was found between the sequences compared. In the 5′ UT region, only one nucleotide difference is observed; it is in the gene for thymidine kinase. (Note that this region is short in most genes and no adequate data are available for many of the genes under study.) In the 3′ UT region, variation is noted in the genes for interleukin-1β, interleukin-5, keratin-18, plasmin inhibitor-α and vasoactive intestinal polypeptide. At the nondegenerate sites, variation is seen in the genes for angiotensinogen, apolipoprotein E, factor IX, fibrinogen-γ, hepatic lipase, phosphoglycerate kinase, plasmin inhibitor-α, tumor antigen p53 and von Willebrand factor. At the twofold degenerate sites, nonsynonymous variation is observed only in the gene for factor VIII while synonymous variation is observed in the genes for apolipoprotein E, fibrinogen-γ, tissue plasminogen activator and von Willebrand factor. At the fourfold degenerate sites, variation is observed in the genes for adenosine deaminase, fibrinogen-γ, hepatic lipase, phosphoglycerate mutase, tissue plasminongen activator, triosephosphate isomerase and von Willebrand factor.

The average level of nucleotide diversity (π) computed from the pooled data is low (Table 1). The highest level is only 0.11%, which is observed at the fourfold degenerate sites. The second highest level is observed at the two-fold degenerate sites; the synonymous and nonsynonymous components are 0.05% and 0.01%, respectively, and the total diversity is 0.06%. The third highest level, $\pi = 0.04\%$, is observed in the 3′ UT regions, though it is not significantly different from the π values in the 5′ UT regions and at the nodegenerate sites, both being 0.03%.

Table 2 shows the proportion of nucleotide differences between sequences that have not been checked carefully against each other. Obviously, the level is much higher than that for carefully checked sequences; for example, it is more than ten times higher in both the 5′ and 3′ UT regions. Therefore, unchecked sequences are not suitable for studying nucleotide diversity.

Deletions and insertions: We have also studied deletions and insertions. Both types of changes are called "gaps" because it is difficult to distinguish between a deletion and an insertion if only two sequences are available. In all the sequences that have been checked carefully, no gap was found in the coding regions in any of the sequences. In the 5′ UT region, a gap has been found in one of the three sequences from the locus for the interleukin-1β gene (Table 3). In the 3′ UT region, a gap has been found in one of the three interleukin-1β sequences and three gaps have been found in the two sequences from the locus coding for the vasoactive intestinal polypeptide. For all the genes studied, the number of gaps per nucleotide site is 0.0002 in both the 5′ and 3′ UT regions (Table 3). These estimates are based on a small number of gaps and so are not reliable. Keeping this caution in mind, we may tentatively conclude that in the 5′ and 3′ UT regions the number of gaps per nucleotide site is about the same as the number of nucleotide differences per site (Table 1).

In the unchecked sequences many gaps were found. For example, in the two sequences from the locus for APRT, we found 7 gaps in the 5′ UT region, one gap in the 3′ UT region, and 42 gaps in introns (Table 3 and footnotes). Among the unchecked sequences the observed number of gaps per nucleotide site is 0.014 in the 5′ UT region and 0.006 in the 3′ UT regions. These values are far too high. Moreover, a gap was found in the coding regions in the sequences from the locus for APRT and another gap was found in the coding regions in the sequences from the locus for 5-lipoxygenase; each of these two gaps would cause a shift in the reading frame of the gene. Obviously, unchecked sequences are not suitable for studying the frequency of gap events in the evolution of nucleotide sequences.

DISCUSSION

Although we have carefully selected sequences that are suitable for the present purpose, the data pre-

OSHIMA, MILLAN and CECENA (1986); 45. KULESH and OSHIMA (1988); 46. HALL *et al.* (1982); 47. HALL *et al.* (1987); 48. TSUJIBO, TIANO and LI (1985); 49. CHUNG, *et al.* (1985); 50. SAKAI *et al.* (1987); 51. TAKENO and LI (1989); 52. MELLENTIN, SMITH and CLEARY (1989); 53. BOEL *et al.* (1984); 54. LEITER, KEUTMANN and GOODMAN (1984); 55. TAKEUCHI and YAMADA (1985); 56. LEITER *et al.* (1985); 57. HENDY *et al.* (1981); 58. VASICEK *et al.* (1983); 59. MICHELSON, MARKHAM and ORKIN (1983); 60. MICHELSON *et al.* (1985); 61. SHANSKE *et al.* (1987); 62. TSUJINO *et al.* (1989); 63. TONE *et al.* (1987); 64. HIROSAWA *et al.* (1988); 65. COOKE *et al.* (1980); 66. TRUONG *et al.* (1984); 67. BECKMAN *et al.* (1985); 68. PLUTZSY *et al.* (1986); 69. IMAI *et al.* (1983); 70. MIYAZAKI *et al.* (1984); 71. RHOADS, DIXIT and ROUFA (1986); 72. WHITE, NEW and DUPONT (1986); 73. SHERMAN *et al.* (1983); 74. LEVANON *et al.* (1985); 75. SUZUKI *et al.* (1987); 76. SHIRAI *et al.* (1988); 77. DEGEN, RAJPUT and REICH (1986); 78. BRADSHAW and DEININGER (1984); 79. FLEMINGTON *et al.* (1987); 80. BROWN *et al.* (1985); 81. MAQUAT, CHILCOTE and RYAN (1985); 82. ZAKUT-HOURI *et al.* (1985); 83. LAMB and CROWFORD (1986); 84. PENNICA *et al.* (1984); 85. NEDWIN *et al.* (1985); 86. TSUKADA *et al.* (1985); 87. MANCUSO *et al.* (1989); 88. SADLER *et al.* (1985); 89. SHELTON-INLOES, TITANI and SADLER (1986); 90. VERWEIJ *et al.* (1986); 91. GINSBURG *et al.* (1989).

[c] The standard errors are computed by assuming binomial sampling of nucleotides.

TABLE 2

Nucleotide differences per site between unchecked sequences

Gene[a]	Noncoding regions		Coding regions				Refs.[b]
				2-fold deg.			
	5'UT	3'	Nondegenerate	Nonsynonymous	Synonymous	4-fold degenerate	
APRT[c]	0/71	0/227	0/329	0/98	0/98	0/110	1, 2
Aldose reductase	2/28	9/372	1/612	0/200	0/200	1/133	3, 4
Butylcholinesterase	1/75	0/485	0/1,177	0/385	0/385	1/241	5, 6
Elongation factor-1α	0/53	1/64	0/911	0/245	0/245	0/227	7, 8
Glutathione peroxidase	1/109	0.7/209	0/381	0/100	0.7/100	0/116	9–11
Hemopoietic cell kinase	NA	7/334	1/987	0/298	0/298	1/227	12, 13
Interleukin-5	0/44	1/367	0/245	0/97	0/97	0/57	14, 15
Keratin 7	0/55	NA	1/895	0/268	1/268	0/241	16, 17
5-Lipoxygenase	3/35	0/428	0/1,309	0/418	0/418	0/289	18, 19
Lysozyme	0.7/10	2/290	0.3/282	0.3/94	0/94	0/64	20–22
Peroxidase (thyroid)	0/72	0/174	3.5/1,780	0.5/523	1/523	2/493	23, 24
Prolyl 4-hydroxylase	3/29	10/401	10/978	0/324	5/324	6/216	25, 26
Thrombomodulin	0/150	2/1,782	1/1,115	0/291	0/291	0/316	27, 28
Total	9.7/730	32.7/5133	17.8/11001	0.8/3341	7.7/3341	11/2730	
	0.0119	0.0064	0.0016	0.0002	0.0023	0.0040	

[a] Three sequences are used for each of the genes for glutathione peroxidase and lysozyme and only two sequences are used for each of the other genes.

[b] 1. BRODERICK *et al.* (1987); 2. HIDAKA *et al.* (1987); 3. BOHREN *et al.* (1989); 4. CHUNG and LA MENDOLA (1989); 5. MCTIERMAN *et al.* (1987); 6. PRODY *et al.* (1987); 7. BRANDS *et al.* (1986); 8. UETSUKI *et al.* (1989); 9. SUKENAGA *et al.* (1987); 10. ISHIDA *et al.* (1987); 11. MULLENBACK *et al.* (1987); 12. QUINTRELL *et al.* (1987); 13. ZIEGLER *et al.* (1987); 14. TANABE *et al.* (1987); 15. CAMPBELL *et al.* (1987); 16. GLASS, KIM and FUCHS (1985); 17. GLASS and FUCHS (1988); 18. DIXON *et al.* (1988); 19. MATSUMOTO *et al.* (1988); 20. CHUNG, KESHAV and GORDON (1988); 21. CASTANON *et al.* (1988); 22. YOSHIMURA, TOIBANA and NAKAHAMA (1988); 23. KIMURA *et al.* (1987); 24. KIMURA *et al.* (1989); 25. PIHLAJANIEMI *et al.* (1987); 26. TASANEN *et al.* (1988); 27. JACKMAN *et al.* (1987); 28. SHIRAI *et al.* (1988).

[c] Adenine phosphoribosyltransferase. Within introns, 12 nucleotide differences in 1664 sites.

sented in Table 1 may not be completely free of errors. This is because in several cases the differences have not been verified. For example, in the case of angiotensinogen, upon seeing a difference between their sequence and that of KAGEYAMA, OHKUBO and NAKANISHI (1984), KUNAPULI and KUMAR (1986) rechecked the accuracy of their own sequence but had not communicated with KAGEYAMA *et al.* to confirm the difference. As another example, in the case of apolipoprotein E, the two sequences were from the same laboratory and represented two alleles (E3 and E4) that had been known to differ by one amino acid, but the authors have made no comment about the three synonymous differences between the two sequences.

The observed nucleotide differences are not distributed randomly among the genes studied, *i.e.*, most of the sequence pairs were identical whereas a number of pairs showed multiple differences. This nonrandom distribution could be partly due to variation in mutation rate among regions (WOLFE, SHARP and LI 1989) and partly due to sequence errors, *i.e.*, because some sequences were obtained with less care than the others. It could also be partly due to differences in sequence lengths and partly due to statistical fluctuations, *e.g.*, the three synonymous differences between the two apolipoprotein E sequences could have occurred because the two alleles have persisted in the population for a long time.

The results in Table 1 may be taken as representing the approximate level of diversity in the American white population, because most of the DNA libraries used were constructed from white Americans. However, it probably represents an upper estimate for two reasons. First, some of the libraries were from Japanese, Europeans, Australians, or other groups. (We have contacted many of the authors for information about the ethnic origins of their sequences but were told by quite a few of them that such information was not available.) Second, as mentioned above, the data used are probably not completely free of sequencing errors.

It should also be noted that the genes used may not represent a random sample of the human genes since they were sequenced because of their biological or medical importance. However, except for the case of apolipoprotein E, all the alleles we compared were wild-type alleles and were not selected for any known mutations or protein variants. We also note that many authors have used the genomic library provided by T. MANIATIS; however, this is unlikely to produce any systematic bias because in each case the individual used for the cDNA library was chosen randomly.

The relative levels of diversity in different DNA regions shown in Table 1 are consistent with the relative degrees of sequence divergence obtained from between species comparisons and, like the later

TABLE 3

Numbers of gaps (deletions or insertions) between sequences in the 5′ and 3′ UT regions

Gene	5′UT	3′UT
A. Checked sequences (same as those in Table 1)		
Interleukin-1β	0.7/77	0.7/599
Vasoactive intestinal polypeptide	0/176	3/767
Other genes	0/3,371	0/18,403
Total	0.7/3,624	3.7/19,769
	0.0002	0.0002
B. Unchecked sequences (same as those in Table 2)		
APRT[a]	7/71	1/227
Aldose reductase	0/28	2/372
Butylcholinesterase	0/75	7/485
Elongation factor-1α	0/53	0/64
Glutathione peroxidase	0/109	0.7/209
Hemopoietic cell kinase	NA	4/334
Interleukin-5[b]	0/44	0/367
Keratin-7[c]	0/55	NA
5-Lipoxygenase[d]	2/35	3/428
Lysozyme	0/10	4/290
Peroxidase (thyroid)	0/71	0/174
Prolyl 4-hydroxylase	0/29	6/401
Thrombomodulin[e]	1/150	4/1,782
Total	10/730	31.7/5,133
	0.0137	0.0062

[a] In the 5′ flanking region, 10 gaps/513 sites; within introns, 42 gaps/1664 sites.

[b] In the 5′ flanking region, 5 gaps/508 sites; within introns, 8 gaps/1265 sites; in the 3′ flanking region, 18 gaps/660 sites.

[c] One single nucleotide deletion (error) in the coding region of the cDNA sequence causing a shift in reading frame.

[d] One two-nucleotide deletion (error) in the coding region of the cDNA sequence causing a shift in the reading frame.

[e] In the 5′ flanking region, 0 gap/402 sites; in the 3′ flanking region, 7 gaps/165 sites.

observations, can be explained by the relative stringencies of selective constraints in different regions (see LI, WU and LUO 1985). For example, in coding regions, nondegenerate sites and fourfold degenerate sites would be subjected to, respectively, the strongest and the weakest selective constraints and would show, respectively, the lowest and the highest level of diversity. In fact, this is the case and the level of diversity at fourfold degenerate sites is about four times higher than that at nondegenerate sites, which is about the same ratio obtained from between species comparisons (see LI, WU and LUO 1985). As another example, the 5′ UT region is in general more conservative than the 3′ UT region and so the diversity at the 5′ UT region would tend to be lower than that at the 3′ UT region, though in the present case the difference is not significant because the sample is not large enough.

Table 1 suggests that the nucleotide diversity in humans is at most of the order of 0.1. There are three possible explanations for this low value: the mutation rate in humans is low, the effective population size of the human species has been relative small in the past, or the species has gone through a severe bottleneck in the recent past. There are two lines of evidence against the second possibility. First, chimpanzees and humans share many common alleles at loci for major histocompatibility complex (*MHC*) genes (LAWLOR *et al.* 1988; MAYER *et al.* 1988). If the human lineage ever went through any severe bottleneck, most of these "trans-species" polymorphisms would have been lost in the human species. Incidentally, it is interesting to note that alleles at a class I *MHC* locus may differ from each other at many nucleotide sites (see MAYER *et al.* 1988). This is in sharp contrast with the observation that most of the sequence pairs in Table 1 are identical while the others differ at most at only a few sites. One simple explanation for the large differences between *MHC* alleles is that they have been maintained in the population for a long time by overdominant selection (HUGHES and NEI 1989). It should be noted that even under overdominant selection many of the polymorphic alleles would have been lost, had a severe bottleneck occurred in the past [see Takahata (1990) for a theoretical treatment]. Second, XIONG *et al.* (1991) estimated that both a Japanese and a Venezuelan apolipoprotein C-II deficiency alleles have persisted in the human population for more than 500,000 years because their nucleotide sequences differ from that of the normal human allele more than does the normal chimpanzee sequence. Had a severe bottleneck occurred in the human population, neither of these alleles would have persisted for so long.

The result in Table 1 can be used to estimate the heterozygosity at the protein level in American whites. By definition, the heterozygosity in a random mating population is the probability that two randomly chosen protein sequences are different. In the present study among the approximately 49 pairs of sequences examined, 10 pairs differ by at least one nonsynonymous difference and so the average heterozygosity at the protein level is 10/49 = 20.4%. Obviously, the heterozygosity for a protein is dependent on its size. The expected heterozygosity for an average protein can be estimated as follows. We first compute the average number of nonsynonymous site per gene (L_N) by counting each nondegenerate site as one nonsynonymous site and each twofold denerate site as 2/3 of a nonsynonymous site. For the 49 genes studied, $L_N = (34{,}869 + 10{,}787 \times 2/3)/49 = 42{,}060/49 \approx 858$. On the other hand, the average number of nucleotide differences per nonsynonymous site is $(10 + 1)/42{,}060 = 0.00026$, since 10 differences were observed at nondegenerate sites and 1 difference was observed at twofold degenerate sites and since the total number of nonsynonymous sites is 42,060. Thus, for an average gene with 858 nonsynonymous sites the probability (heterozygosity) that two randomly chosen sequences differ by at least one nonsynonymous difference is $1 - (1 - 0.00026)^{858} = 20.0\%$,

TABLE 4

Nucleotide diversity in species of *Drosophila*

Regions[a]	*D. pseudoobscura* Length (kb)	*D. pseudoobscura* π	*D. simulans* Length (kb)	*D. simulans* π	*D. melanogaster* Length (kb)	*D. melanogaster* π
Adh	32	0.026	13	0.015	13	0.006
Amy	26	0.019			15	0.008
rosy	5	0.013	100	0.018	100	0.005
Average (weighted)		0.022		0.018		0.005

Data compiled by AQUADRO (1991). Sample sources: in *D. pseudoobscura*, *Adh* from California (1 location); *Amy* from California (3 locations), British Columbia, Baja California (Mexico), Hidalgo (Mexico), and Bogota (Columbia); *rosy* from several lines from several locations across the western United States. In *D. simulans* and *D. melanogaster*, *adh* and *Amy* from several east coast United States populations; *rosy* from a single collection site.

[a] Only autosomal regions are used so that a comparison can be made with the nucleotide diversity in Table 1, which is estimated mostly from autosomal genes.

which is close to the average heterozygosity computed above.

The above computation refers to the heterozygosity at the protein sequence level. At the electrophoretic level, the heterozygosity is expected to be lower. NEI and CHAKRABORTY (1973) estimated that about 33% of amino acid changes are expected to cause a charge change in a protein. Using this value, we estimate that for an average protein the expected heterozygosity at the electrophoretic level is $1-(1-0.33\times 0.00026)^{858} \approx 7.4\%$. This is close to the average heterozygosity (7.3%) in whites of European origin computed from electrophoretic data of 87 loci (HARRIS, HOPKINSON and EDWARDS 1977).

We now compare the nucleotide diversity in man with those in Drosophila populations. Table 4 represents a summary of the nucleotide diversity in three *Drosophila* species estimated from restriction enzyme data. Although restriction enzymes do not detect all variation between sequences, the level of nucleotide diversity estimated by this method for the Adh region (LANGLEY, MONTGOMERY and QUATTLEBAUM 1982) is the same as that obtained by direct nucleotide sequencing (KREITMAN 1983). Moreover, results of detailed surveys of the *rosy* region by 4-base recognition enzymes in both *D. pseudoobscura* (RILEY, HALLAS and LEWONTIN 1989) and *Drosophila melanogaster* (C. F. AQUADRO, personal communication) gave similar estimates as those obtained from 6-base recognition enzymes. We may therefore assume that the values in Table 4 are reasonably accurate estimates of the levels of nucleotide diversity in the regions studied. The average of the estimates in each species may be taken as the average nucleotide diversity in noncoding regions because the three regions contain mostly sequences that do not code for amino acids.

In *D. pseudoobscura* the average nucleotide diversity for the three regions is 2.2% (Table 4). Since this represents largely the diversity in noncoding regions, it cannot be directly compared with the estimates for humans shown in Table 1. However, RILEY, HALLAS and LEWONTIN (1989) and M. RILEY (personal communication) have found that the silent sites in the *XDH* have a higher nucleotide diversity than introns and the 5′ flanking region. Moreover, comparative analyses have shown that in mammalian genes the rate of nucleotide substitution at fourfold degenerate sites is only slightly lower than that in pseudogenes, suggesting that fourfold degenerate sites are subject to only weak selective constraints (see LI and GRAUR 1991). Therefore, the nucleotide diversity at fourfold degenerate sites in human genes ($\pi = 0.11\%$) is unlikely to be twofold lower than that in noncoding regions. Although the human genes studied were mainly from American whites, the Drosophila samples were from even smaller regions in America (see Table 4 footnotes). We may therefore conclude that in noncoding regions the nucleotide diversity in humans is one order of magnitude lower that in *D. pseudoobscura* ($\pi = 2.2\%$). This is in sharp contrast to the observation that at the electrophoretic level, humans and *D. pseudoobscura* show similar levels of genetic variability (see the Introduction).

Why should the level of nucleotide diversity differ so much between the two species, though at the electrophoretic level the average heterozygosity is similar in the two species? A simple explanation is to assume: (1) the majority of nucleotide changes in nonconding regions are selectively neutral or almost neutral whereas the majority of electrophoretic variants are slightly deleterious and (2) the long-term effective population size in *D. pseudoobscura* is considerably (say, ten times) larger than that in *H. sapiens*. Under these two assumptions the level of nucleotide diversity will be much higher in *D. pseudoobscura* than in *H. sapiens* because a larger population can accumulate more neutral mutations than a smaller one. On the other hand, the former species may not necessarily have a higher heterozygosity for electrophoretic variants than the latter because selection against slightly deleterious mutations is more effective in a large population than in a small one. If the above two assumptions are correct, then the contrast in the extent of genetic variability at the two levels between the two species supports OHTA's (1974) hypothesis of slightly deleterious mutation. A similar explanation for the lower nucleotide diversity in *D. melanogaster* than in *D. simulans* has been put forward earlier by AQUADRO, LADO and NOON (1988). For a different view on the maintenance of nucleotide diversity in Drosophila, see KREITMAN and HUDSON (1991).

The average nucleotide diversity in *D. melanogaster* is 0.05% (Table 4). This is much lower than that in

D. pseudoobscura, probably due to a smaller effective population size (CHOUDHARY and SINGH 1987; AQUADRO 1991). However, it is considerably higher than that in humans. Since among all the *Drosophila* species studied to date, *D. melanogaster* has the lowest level of nucleotide diversity (see the review by AQUADRO 1991), we may conclude that the level of nucleotide diversity in human populations is much lower than those in Drosophila populations.

Finally, we discuss how to test the significance of difference in nucleotide diversity (π) between two populations or species. For this purpose one first computes the variance of π for each population. For nucleotide sequence data, this variance can be computed by assuming binomial sampling of nucleotides as in Table 1. That is, $V(\pi) = \pi(1 - \pi)/L$, where L is the total number of nucleotide sites compared and π is the average value over all sequences. For restriction enzyme data, the variance can be computed by using NEI and TAJIMA's (1981) formula. Now suppose that we have two values, π_1 and π_2, and want to test the hypothesis $\pi_1 = \pi_2$. Let $d = \pi_1 - \pi_2$. Then the variance of d is given by $V(d) = V(\pi_1) + V(\pi_1)$ and the standard error $\sigma(d)$ is the square root of $V(d)$. Under the assumption of normal distribution, one can test whether the mean of d is larger than $2\sigma(d)$. The assumption of normality should hold approximately if both L_1 and L_2 are large. If either L_1 or L_2 is small, one may use the t-test; however, one problem with the t-test is that the assumption of equal variances, *i.e.*, $V(\pi_1) = V(\pi_2)$, may not hold well. Since the above hypothesis is $\pi_1 = \pi_2$, we use a two-sided test. If the hypothesis is, say, $\pi_1 > \pi_2$, then a one-sided test should be used.

In the preceding discussion we have not considered the effect of sequencing errors. This factor is not easy to treat because it depends on the skill and carefulness of the person who does the sequencing work and also depends on other factors such as the GC-richness of the sequences. However, if the error rate is known to be at least e, then to test whether d is significantly different from 0, one may test whether $|d| - e$ is significantly greater than 0. On the other hand, for testing the hypothesis of $\pi_1 > \pi_2$, one may test whether $\pi_1 - \pi_2 - e$ is significantly greater than 0.

In the above comparison of the π values between humans and *Drosophila* species we have not conducted a formal statistical test because the variance of π is not available for the latter. However, the difference in π values is so large that the difference is probably statistically significant.

We thank C. F. AQUADRO and M. KREITMAN for sending us preprints and M. RILEY, PAUL SHARP and a reviewer for suggestions. This study was supported by U.S. Public Health Servie grants.

LITERATURE CITED

ADRIAN, G. S., D. A. WIGINTON and J. J. HUTTON, 1984 Structure of adenosine deaminase mRNAs from normal and Adenosine deaminase-deficient human cell line. Mol. Cell. Biol. **4:** 1712–1717.

AQUADRO, C. F., 1991 Molecular population genetics of *Drosophila*, in *Molecular Approaches to Pure and Applied Entomology*, edited by J. OAKESHOTL and M. WHITTEN. Springer-Verlag, New York (in press).

AQUADRO, C. F., K. M. LADO and W. A. NOON, 1988 The rosy region of *Drosophila melanogaster* and *Drosophila simulans*. I. Contrasting levels of naturally occurring DNA restriction map variation and divergence. Genetics **119:** 875–888.

AZUMA, C., T. TANABE, M. KONISHI, T. KINASHI, T. NOMA, F. MATSUDA, Y. YAOITA, K. TAKATSU, L. HAMMARSTRÖM, C. I. E. SMITH, E. SEVERINSON and T. HONJO, 1986 Cloning of cDNA for human T-cell replacing factor (interleukin-5) and comparison with the murine homologue. Nucleic Acids Res. **14:** 9149–9158.

BECKMANN, R. J., R. J. SCHMIDT, R. F. SANTERRE, J. PLUTZKY, G. CRABTREE and G. L. LONG, 1985 The structure and evolution of a 461 amino acid human protein C precursor and its messenger RNA, based upon the DNA sequence of cloned human liver cDNAs. Nucleic Acids Res. **13:** 5233–5247.

BENSI, G., R. RAUGEI, E. PALLA, V. CARINCI, D. T. BUONAMASSA and M. MELLI, 1987 Human interleukin-1 beta gene. Gene **52:** 95–101.

BOEL, E., T. W. SCHWARTZ, K. E. NORRIS and N. P. FIIL, 1984 A cDNA encoding a small common precursor for human pancreative polypeptide and pancreative icosapeptide. EMBO J. **3:** 909–912.

BOHREN, K. M., B. BULLOCK, B. WERMUTH and K. H. GABBAY, 1989 The aldo-keto reductase superfamily: cDNA and deduced amino acid sequences of human aldehyde and aldose reductases. J. Biol. Chem. **264:** 9547–9551.

BRADSHAW, H. D., JR., and P. L. DEININGER, 1984 Human thymidine kinase gene: molecular cloning and nucleotide sequence of a cDNA expressible in mammalian cells. Mol. Cell. Biol. **4:** 2316–2320.

BRANDS, J. H. G. M., J. A. MAASSEN, F. J. VAN HEMERT, R. AMONS and W. MÖLLER, 1986 The primary structure of the subunit of human elongation factor 1 structural aspects of guanine-nucleotide-binding sites. Eur. J. Biochem. **155:** 167–171.

BRODERICK, T. P., D. A. SCHAFF, A. M. BERTINO, M. K. DUSH, J. A. TISCHFIELD and P. J. STAMBROOK, 1987 Comparative anatomy of the human APRT gene and enzyme: nucleotide sequence divergence and conservation of a nonrandom CpG dinucleotide arrangement. Proc. Natl. Acad. Sci. USA **84:** 3349–3353.

BROWN, J. R., I. O. DAAR, J. R. KRUG and L. E. MAQUAT, 1985 Characterization of the functional gene and several processed pseudogenes in the human triosephosphate isomerase gene family. Mol. Cell. Biol. **5:** 1694–1706.

CAI, S.-J., D. M. WONG, S.-H. CHEN and L. CHAN, 1989 Structure of the human hepatic triglyceride lipase gene. Cell Biol. **28:** 8966–8971.

CAMPBELL, H. D., W. Q. J. TUCKER, Y. HORT, M. E. MARTINSON, G. MAYO, E. J. CLUTTERBUCK, C. J. SANDERSON and I. G. YOUNG, 1987 Molecular cloning, nucleotide sequence, and expression of the gene encoding human eosinophil differentiation factor (interleukin 5). Proc. Natl. Acad. Sci. USA **84:** 6629–6633.

CARTER, P. E., B. DUNBAR and J. E. FOTHERGILL, 1988 Genomic abd cDNA cloning of the human C1 inhibitor intron-exon junctions and comparison with other serpins. Eur. J. Biochem. **173:** 163–169.

CASTANON, M. J., W. SPEVAK, G. R. ADOLF, E. CHLEBOWICZ-

SLEDZIEWSKA and A. SLEDZIEWSKI, 1988 Cloning of human lysozyme gene and expression in the yeast *Saccharomyces cerevisiae*. Gene **66:** 223–234.

CHOUDHARY, M., and R. S. SINGH, 1987 Historical effective size and the level of genetic diversity in *Drosophila melanogaster* and *Drosophila pseudoobscura*. Biochem. Genet. **25:** 41–51.

CHUNG, D. W., and E. W. DAVIE, 1984 γ and γ' chains of human fibrinogen are produced by alternative mRNA processing. Biochemistry **23:** 4232–4236.

CHUNG, F.-Z., H. TSUJIBO, U. BHATTACHARYYA, F. S. SHARIEF and S. S.-L. LI, 1985 Genomic organization of human lactate dehydrogenase-A gene. Biochem. J. **231:** 537–541.

CHUNG, L. P., S. KESHAV and S. GORDON, 1988 Cloning the human lysozyme cDNA: inverted ALu repeat in the mRNA and *in situ* hybridization for macrophages and paneth cells. Proc. Natl. Acad. Sci. USA **85:** 6227–6231.

CHUNG, S., and J. LAMENDOLA, 1989 Cloning and sequence determination of human placental aldose reductase gene. J. Biol. Chem. **264:** 14775–14777.

CLARK, B. D., K. L. COLLINS, M. S. GANDY, A. C. WEBB and P. E. AURON, 1986 Genomic sequence for human prointerleukin 1 beta: Possible evolution from a reverse transcribed prointerleukin 1 alpha gene. Nucleic Acids Res. **14:** 7897 -7914.

COOKE, N. E., D. COIT, R. I. WEINER, J. D. BAXTER and J. A. MARTIAL, 1980 Structure of cloned DNA complementary to rat prolactin messenger RNA. J. Biol. Chem. **255:** 6502–6510.

COSTANZO, F., C. SANTORO, V. COLANTUONI, G. BENSI, G. RAUGEI, V. ROMANO and R. CORTESE, 1984 Cloning and sequencing of a full length cDNA coding for a human apoferritin H chain: evidence for a multigene family. EMBO J. **3:** 23–27.

COSTANZO, F., M. COLOMBO, S. STAEMPFLI, C. SANTORO, M. MARONE, R. FRANK, H. DELIUS and R. CORTESE, 1986 Structure of gene and pseudogenes of human apoferritin H. Nucleic Acids Res. **14:** 721–736.

DATTA, S., C.-C. LUO, W.-H. LI, P. VANTUINEN, D. H. LEDBETTER, M. A. BROWN, S.-H. CHEN, S.-W. LIU and L. CHAN, 1988 Human hepatic lipase: cloned cDNA sequence, restriction fragment length polymorphisms, chromosomal localization, and evolutionary relationships with lipoprotein lipase and pancreatic lipase. J. Biol. Chem. **263:** 1107–1110.

DEGEN, S. J. F., B. RAJPUT and E. REICH, 1986 The human tissue plasminogen activator gene. J. Biol. Chem. **261:** 6972–6985.

DENTE, L., G. CILIBERTO and R. CORTESE, 1985 Structure of the human α_1-acid glycoprotein gene: sequence homology with other human acute phase protein genes. Nucleic Acids Res. **13:** 3941–3952.

DIXON, R. A. F., R. E. JONES, R. E. DIEHL, C. D. BENNETT, S. KARGMAN and C. A. ROUZER, 1988 Cloning of the cDNA for human 5-lipoxygenase. Proc. Natl. Acad. Sci. USA **85:** 416–420.

FLEMINGTON, E., H. D. BRADSHAW, JR., V. TRAINA-DORGE, V. SLAGEL and P. L. DEININGER, 1987 Sequence, structure and promoter characterization of the human thymidine kinase gene. Gene **52:** 267–277.

GINSBURG, D., B. A. KONKLE, J. C. GILL, R. R. MONTGOMERY, P. L. BOCKENSTEDT, T. A. JOHNSON and A. Y. YANG, 1989 Molecular basis of human von Willebrand disease: analysis of platelet von Willebrand factor mRNA. Proc. Natl. Acad. Sci. USA **86:** 3723–3727.

GITSCHIER, J., W. I. WOOD, T. M. GORALKA, K. L. WION, E. Y. CHEN, D. H. EATON, G. A. VEHAR, D. J. CAPON and R. M. LAWN, 1984 Characterization of the human factor VIII gene. Nature **312:** 326–330.

GLASS, C., and E. FUCHS, 1988 Isolation, sequence, and differential expression of human K7 gene in simple epithelial cells. J. Cell. Biol. **107:** 1337–1350.

GLASS, C., K. H. KIM and E. FUCHS, 1985 Sequence and expression of a human type II mesothelial keratin. J. Cell Biol. **101:** 2366–2373.

GRAHAM, A., P. J. HEDGE, S. J. POWELL, J. RILEY, L. BROWN, A. GAMMACK, F. CAREY and A. F. MARKHAM, 1989 Nucleotide sequence of cDNA for human aldose reductase. Nucleic Acids Res. **17:** 8368–8368.

HALL, L., R. K. CRAIG, M. R. EDBROOKE and P. N. CAMPBELL, 1982 Comparison of the nucleotide sequence of cloned human and guinea-pig pre-alpha-lactalbumin cDNA with that of chick pre-lysozyme cDNA suggests evolution from a common ancestral gene. Nucleic Acids Res. **10:** 3503–3515.

HALL, L., D. C. EMERY, M. S. DAVIES, D. PARKER and R. K. CRAIG, 1987 Organization and sequence of the human alpha-lactalbumin gene. Biochem. J. **242:** 735–742.

HARRIS, H., 1966 Enzyme polymorphisms in man. Proc. R. Soc. Lond. Ser. B **164:** 198–310.

HARRIS, H., D. A. HOPKINSON and Y. H. EDWARDS, 1977 Polymorphism and the subunit structure of enzymes: a contribution to the neutralist-selectionist controversy. Proc. Natl. Acad. Sci. USA **74:** 698–701.

HENDY, G. N., H. M. KRONENBERG, J. T. POTTS, JR., and A. RICH, 1981 Nucleotide sequence of cloned cDNAs encoding human preproparathyroid hormone. Proc. Natl. Acad. Sci. USA **78:** 7365–7369.

HIDAKA, Y., TARLE, S. A, T. E. O. TOOLE, W. N. KELLEY and T. D. PALELLA, 1987 Nucleotide sequence of the human APRT gene. Nucleic Acids Res. **15:** 9086–9086.

HIROSAWA, S., Y. NAKAMURA, O. MIURA, Y. SUMI and N. AOKI, 1988 Organization of the human alpha-2-plasmin inhibitor gene. Proc. Natl. Acad. Sci. USA **85:** 6836–6840.

HOHN, P. A., N. C. POPESCU, R. D. HANSON, G. SALVESEN and T. J. LEY, 1989 Genomic organization and chromosomal localization of the human cathepsin G gene. J. Biol. Chem. **264:** 13412–13419.

HUGHES, A. L., and M. NEI, 1989 Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc. Natl. Acad. Sci. USA **86:** 958–962.

IMAI, T., H. MIYAZAKI, S. HIROSE, H. HORI, T. HAYASHI, R. KAGEYAMA, H. OHKUBO, S. NAKANISHI and K. MURAKAMI, 1983 Cloning and sequence analysis of cDNA for human renin precursor. Proc. Natl. Acad. Sci. USA **80:** 7405–7409.

ISHIDA, K., T. MORINO, K. TAKAGI and Y. SUKENAGA, 1987 Nucleotide sequence of a human gene for glutathione peroxidase. Nucleic Acids Res. **15:** 10051–10051.

JACKMAN, R. W., D. L. BEELER, L. FRITZE, G. SOFF and R. D. ROSENBERG, 1987 Human thrombomodulin gene is intron depleted: nucleic acid sequences of the cDNA and gene predict protein structure and suggest sites of regulatory control. Proc. Natl. Acad. Sci. USA **84:** 6425–6429.

JACOBS, K., C. SHOEMAKER, R. RUDERSDORF, S. D. NEILL, R. J. KAUFMAN, A. MUFSON, J. SEEHRA, S. S. JONES, R. HEWICK, E. F. FRITSH, M. KAWAKITA, T. SHIMIZU and T. MIYATA, 1985 Isolation and characterization of genomic and cDNA clones of human erythropoietin. Nature **313:** 806–810.

JONAS, V., C. R. LIN, E. KAWASHIMA, D. SEMON, L. W. SWANSON, J.-J. MERMOD, R. M. EVANS and M. G. ROSENFELD, 1985 Alternative RNA processing events in human calcitonin/calcitonin gene-related peptide gene expression. Proc. Natl. Acad. Sci. USA **82:** 1994–1998.

KAGEYAMA, R., H. OHKUBO and S. NAKANISHI, 1984 Primary structure of human preangiotensinogen deduced from the cloned cDNA sequence. Biochemistry **23:** 3603–3609.

KIMURA, S., T. KOTANI, O. W. MCBRIDE, K. UMEKI, K. HIRAI, T. NAKAYAMA and S. OHTAKI, 1987 Human thyroid peroxidase: complete cDNA and protein sequence, chromosome mapping, and identification of two alternately spliced mRNAs. Proc. Natl. Acad. Sci. USA **84:** 5555–5559.

KIMURA, S., Y.-S. HONG, T. KOTANI, S. OHTAKI and F. KIKKAWA,

1989 Structure of the human thyroid peroxidase gene: comparison and relationship to the human myeloperoxidase gene. Biochemistry **28:** 4481–4489.

KREITMAN, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature **304:** 412–417.

KREITMAN, M., and R. R. HUDSON, 1991 Inferring the evolutionary histories of the Adh and Adh-dup loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics **127:** 565–582.

KULESH, D. A., and R. G. OSHIMA, 1988 Cloning of the human keratin 18 gene and its expression in nonepithelial mouse cells. Mol. Cell. Biol. **8:** 1540–1550.

KUNAPULI, S. P., and A. KUMAR, 1986 Difference in the nucleotide sequence of human angiotensinogen cDNA. Nucleic Acids Res. **14:** 7509–7509.

KURACHI, K., E. W. DAVIE, D. J. STRYDOM, J. F. RIORDAN and B. L. VALLEE, 1985 Sequence of the cDNA and gene for angiogenin, a human angiogenesis factor. Biochemistry **24:** 5494–5499.

LAMB, P., and L. CRAWFORD, 1986 Characterization of the human p53 gene. Mol. Cell. Biol. **6:** 1379–1385.

LANGLEY, C. H., E. MONTGOMERY and W. F. QUATTLEBAUM, 1982 Restriction map variation in the *Adh* region of *Drosophila*. Proc. Natl. Acad. Sci. USA **79:** 5631–5635.

LAWLOR, D. A., F. E. WARD, P. D. ENNIS, A. P. JACKSON and P. PARHAM, 1988 HLA-A and B polymorphisms predate the divergence of humans and chimpanzees. Nature **335:** 268–271.

LEITER, A. B., H. T. KEUTMANN and R. H. GOODMAN, 1984 Structure of a precursor to human pancreative polypeptide. J. Biol. Chem. **259:** 14702–14705.

LEITER, A. B., M. R. MONTMINY, E. JAMIESON and R. H. GOODMAN, 1985 Exons of the human pancreatic polypeptide gene define functional domains of the precursor. J. Biol. Chem. **260:** 13013–13017.

LEVANON, D., J. LIEMAN-HURWITZ, N. DAFNI, M. WIGDERSON, L. SHERMAN, Y. BERNSTEIN, Z. LAVER-RUDICH, E. DANCIGER, O. STEIN and Y. GRONER, 1985 Architecture and anatomy of the chromosomal locus in human chromosome 21 encoding the Cu/Zn superoxide dismutase. EMBO J. **4:** 77–84.

LEWONTIN, R. C., 1974 *The Genetic Basis of Evolutionary Change.* Columbia University Press, New York.

LEYTUS, S. P., D. C. FOSTER, K. KURACHI and E. W. DAVIE, 1986 Gene for human factor X: a blood coagulation factor whose gene organization is essentially identical with that of factor IX and protein C. Biochemistry **25:** 5098–5102.

LI, W.-H., and D. GRAUR, 1991 *Fundamentals of Molecular Evolution.* Sinauer Associates, Sunderland, Mass.

LI, W.-H., C.-I. WU and C.-C. LUO, 1985 A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2:** 150 -174.

LUBAHN, D. B., D. R. JOSEPH, M. SAR, J. TAN, H. N. HIGGS, R. E. LARSON, F. S. FRENCH and E. M. WILSON, 1988 The human androgen receptor: complementary deoxyribonucleic acid cloning, sequence analysis and gene expression in prostate. Mol. Endocrinol. **2:** 1269–1275.

LUBAHN, D. B., T. R. BROWN, J. A. SIMENTAL, H. N. HIGGS, C. J. MIGEON, E. M. WILSON and F. S. FRENCH, 1989 Sequence of the intron/exon junctions of the coding region of the human androgen receptor gene and identification of a point mutation in a family with complete androgen insensitivity. Proc. Natl. Acad. Sci. USA **86:** 9534–9538.

MANCUSO, D. J., E. A. TULEY, L. A. WESTFIELD, N. K. WORRALL, B. B. SHELTON-INLOES, J. M. SORACE, Y. G. ALEVY and J. E. SADLER, 1989 Structure of the gene for human von Willebrand facotr. J. Biol. Chem. **264:** 19514–19527.

MAQUAT, L. E., R. CHILCOTE and P. M. RYAN, 1985 Human triosephosphate isomerase cDNA and protein structure: studies of triosephosphate isomerase deficiency in man. J. Biol. Chem. **260:** 3748–3753.

MATSUMOTO, T., C. D. FUNK, O. RAADMARK, J.-O. HOEOEG, H. JOERVALL and B. SAMUELSSON, 1988 Molecular cloning and amino acid sequence of human 5-lipoxygenase. Proc. Natl. Acad. Sci. USA **85:** 26–30.

MAYER, W. E., M. JONKER, D. KLEIN, P. IVANYI, G. VAN SEVENTER and J. KLEIN, 1988 Nucleotide sequences of chimpanzee MHC class I alleles: evidence for trans-species mode of evolution. EMBO J. **7:** 2765–2774.

MCGRAW, R. A., L. M. DAVIS, C. M. NOYES, R. L. LUNDBLAD, H. R. ROBERTS, J. B. GRAHAM and D. W. ATAFFORD, 1985 Evidence for a prevalent dimorphism in the activation peptide of human coagulation factor IX. Proc. Natl. Acad. Sci. USA **82:** 2847–2851.

MCLEAN, J. W., N. A. ELSHOURBAGY, D. J. CHANG, R. W. MAHLEY and J. M. TAYLOR, 1984 Human apolipoprotein E mRNA: cDNA cloning and nucleotide sequencing of a new variant. J. Biol. Chem. **259:** 6498–6504.

MCTIERNAN, C., S. ADKINS, A. CHATONNET, T. A. VAUGHAN, C. F. BARTELS, M. KOTT, T. L. ROSENBERRY, B. N. LA DU and O. LOCKRIDGE, 1987 Brain cDNA clone for human cholinesterase. Proc. Natl. Acad. Sci. USA **84:** 6682–6686.

MELLENTIN, J. D., S. D. SMITH and M. L. CLEARY, 1989 Lyl-1, a novel gene altered by chromosomal translocation in T cell leukemia, codes for a protein with a helix-loop-helix DNA binding motif. Cell **58:** 77–83.

MICHELSON, A. M., A. F. MARKHAM and S. H. ORKIN, 1983 Isolation and DNA sequence of a full-length cDNA clone for human X chromosome-encoded phosphoglycerate kinase. Proc. Natl. Acad. Sci. USA **80:** 472–476.

MICHELSON, A. M., C. C. F. BLAKE, S. T. EVANS and S. H. ORKIN, 1985 Structure of the human phosphoglycerate kinase gene and the intron-mediated evolution and dispersal of the nucleotide-binding domain. Proc. Natl. Acad. Sci. USA **82:** 6965–6969.

MIYAZAKI, H., A. FUKAMIZU, S. HIROSE, T. HAYASHI, H. HORI, H. OHKUBO, S. NAKANISHI and K. MURAKAMI, 1984 Structure of the human renin gene. Proc. Natl. Acad. Sci. USA **81:** 5999–6003.

MOORE, M. N., F. T. KAO, Y. K. TSAO and L. CHAN, 1984 Human apolipoprotein A-II: nucleotide sequence of a cloned cDna, and localization of its structural gene on human chromosome 1. Biochem. Biophys. Res. Commun. **123:** 1–7.

MULLENBACH, G. T., A. TABRIZI, B. D. IRVINE, G. I. BELL and R. A. HALLEWELL, 1987 Sequence of a cDNA coding for human glutathione peroxidase confirms TGA encodes active site selenocysteine. Nucleic Acids Res. **15:** 5484–5484.

NAGATA, S., M. TSUCHIYA, S. ASANO, Y. KAZIRO, T. YAMAZAKI, O. YAMAMOTO, Y. HIRATA, N. KUBOTA, M. OHEDA, H. NOMURA and M. ONO, 1986a Molecular cloning and expression of cDNA for human granulocyte colony-stimulating factor. Nature **319:** 415–418.

NAGATA, S., M. TSUCHIYA, S. ASANO, O. YAMAMOTO, Y. HIRATA, N. KUBOTA, M. OHEDA, H. NOMURA and T. YAMAZAKI, 1986b The chromosomal gene structure and two mRNAs for human granulocyte colony-stimulating factor. EMBO J. **5:** 575 -581.

NAKAMURA, Y., M. OGAWA, T. NISHIDE, M. EMI, G. KOSAKI, S. HIMENO and K. MATSUBARA, 1984 Sequences of cDNAs for human salivary and pancreatic α-amylases. Gene **28:** 263–270.

NEDWIN, G. E., S. L. NAYLOR, A. L. SAKAGUCHI, D. SMITH, J. JARRETT-NEDWIN, J. V. PENNICA, D. GOEDDEL and P. W. GRAY, 1985 Human lymphotoxin and tumor necrosis factor genes: structure, homology and chromosomal localization. Nucleic Acids Res. **13:** 6361–6373.

NEI, M., and R. CHAKRABORTY, 1973 Genetic distance and elec-

trophoretic identity of proteins between taxa. J. Mol. Evol. **2:** 323–328.

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

NEI, M., and W.-H. LI, 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA **76:** 5269–5273.

NEI, M., and F. TAJIMA, 1981 DNA polymorphism detectable by restriction endonucleases. Genetics **97:** 145–163.

NISHIDA, T., N. NISHINO, M. TAKANO, K. KAWAI, K. BANDO, Y. MASUI, S. NAKAI and Y. HIRAI, 1987 cDNA cloning of IL-1-alpha and IL-1 beta from mRNA of U937 cell line. Biochem. Biophys. Res. Commun. **143:** 345–352.

NISHIDE, T., Y. NAKAMURA, M. EMI, T. YAMAMOTO, M. OGAWA, T. MORI and K. MATSUBARA, 1986 Primary structure of human salivary α-amylase gene. Gene **41:** 299–304.

OHTA, T. , 1974 Mutational pressure as the main cause of molecular evolution and polymorphism. Nature **252:** 351–354.

OSHIMA, R. G., J. L. MILLAN and G. CECENA, 1986 Comparison of mouse and human keratin 18: a component of intermediate filaments expressed prior to implantation. Differentiation **33:** 61–68.

PAIK, Y.-K., D. J. CHANG, C. A. REARDON, G. E. DAVIES, R. W. MAHLEY and J. M. TAYLOR, 1985 Nucleotide sequence and structure of the human apolipoprotein E gene. Proc. Natl. Acad. Sci. USA **82:** 3445–3449.

PENNICA, D., G. E. NEDWIN, J. S. HAYFLICK, P. H. SEEBURG, R. DERYNCK, M. A. PALLADINO, W. J. KOHR, B. B. AGGARWAL and D. V. GOEDDEL, 1984 Human tumor necrosis factor: precursor structure, expression and homology to lymphotoxin. Nature **312:** 724–729.

PIHLAJANIEMI, T., T. HELAAKOSKI, K. TASANEN, R. MYLLYLAE, M.-L. HUHTALA, J. KOIVU and K. I. KIVIRIKKO, 1987 Molecular cloning of the β-subunit of human prolyl 4-hydroxylase. This subunit and protein disulphide isomerase are products of the same gene. EMBO J. **6:** 643–649.

PLUTZKY, J., J. A. HOSKINS, G. L. LONG and G. R. CRABTREE, 1986 Evolution and organization of the human protein C gene. Proc. Natl. Acad. Sci. USA **83:** 546–550.

PRODY, C. A., D. ZEVIN-SONKIN, A. GNATT, O. GOLDBERG and H. SOREQ, 1987 Isolation and characterization of full-length cDNA clones coding for cholinesterase from fetal human tissues. Proc. Natl. Acad. Sci USA **84:** 3555–3559.

QUINTRELL, N., R. LEBO, H. VARMUS, J. M. BISHOP, M. J. PETTENATI, M. M. LE BEAU, M. O. DIAZ and J. D. ROWLEY, 1987 Identification of a human gene (HCK) that encodes a protein-tyrosine kinase and is expressed in hemopoietic cells. Mol. Cell. Biol. **7:** 2267–2275.

RHOADS, D. D., A. DIXIT and D. J. ROUFA, 1986 Primary structure of human ribosomal protein S14 and the gene that encodes it. Mol. Cell. Biol. **6:** 2774 -2783.

RILEY, M. A., M. E. HALLAS and R. C. LEWONTIN, 1989 Distinguishing the forces controlling genetic variation at the *Xdh* locus in *Drosophila pseudoobscura*. Genetics **123:** 359–369.

RIXON, M. W., D. W. CHUNG and E. W. DAVIE, 1985 Nucleotide sequence of the gene for the chain of human fibrinogen. Biochemistry **24:** 2077–2086.

SADLER, J. E., B. B. SHELTON-INLOES, J. M. SORACE, J. M. HARLAN, K. TITANI and E. W. DAVIE, 1985 Cloning and characterization of two cDNAs coding for human von Willebrand factor. Proc. Natl. Acad. Sci. USA **82:** 6394–6398.

SAKAI, K., F. S. SHARIEF, Y.-C. E. PAN and S. S.-L. LI, 1987 The cDNA and protein sequences of human lactate dehydrogenase B. Biochem. J. **248:** 933–936.

SALVESEN, G., D. FARLEY, J. SHUMAN, A. PRZYBYLA, C. REILLY and J. TRAVIS, 1987 Molecular cloning of human cathepsin G: structural similarity to mast cell and cytotoxic T lymphocyte proteinases. Biochemistry **26:** 2289–2293.

SEILHAMER, J. J., A. A. PROTTER, P. FROSSARD and B. LEVY-WILSON, 1984 Isolation and DNA sequence of full-length cDNA and of the entire gene for human apolipoprotein AI-discovery of a new genetic polymorphism in the apo AI gene. DNA **3:** 309–317.

SHANSKE, S., S. SAKODA, M. A. HERMODSON, S. DIMAURO and E. A. SCHON, 1987 Isolation of a cDNA encoding the muscle-specific subunit of human phosphoglycerate mutase. J. Biol. Chem. **262:** 14612–14617.

SHELTON-INLOES, B. B., K. TITANI and J. E. SADLER, 1986 cDNA sequences for human von Willebrand factor reveal five types of repeated domains and five possible protein sequence polymorphisms. Biochemistry **25:** 3164–3171.

SHERMAN, L., N. DAFNI, J. LIEMAN-HURWITZ and Y. GRONER, 1983 Nucleotide sequence and expression of human chromosome 21-encoded superoxide dismutase mRNA. Proc. Natl. Acad. Sci. USA **80:** 5465–5469.

SHIRAI, T., S. SHIOJIRI, H. ITO, S. YAMAMOTO, H. KUSUMOTO, Y. DEYASHIKI, I. MARUYAMA and K. SUZUKI, 1988 Gene structure of human thrombomodulin, a cofactor for thrombin-catalyzed activation of protein C. J. Biochem. **103:** 281 -285.

SUKENAGA, Y., K. ISHIDA, T. TAKEDA and K. TAKAGI, 1987 cDNA sequence coding for human glutathione peroxidase. Nucleic Acids Res. **15:** 7178–7178.

SUZUKI, K., H. KUSUMOTO, Y. DEYASHIKI, J. NISHIOKA, I. MARUYAMA, M. ZUSHI, S. KAWAHARA, G. HONDA, S. YAMAMOTO and S. HORIGUCHI, 1987 Structure and expression of human thrombomodulin, a thrombin receptor on endothelium acting as a cofactor for protein C activation. EMBO J. **6:** 1891–1897.

TAKAHATA, N. , 1990 A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. Proc. Natl. Acad. Sci. USA **87:** 2419–2423.

TAKENO, T., and S. S.-L. LI, 1989 Structure of the human lactate dehydrogenase B Gene. Biochem. J. **257:** 921–924.

TAKEUCHI, T., and T. YAMADA, 1985 Isolation of a cDNA clone encoding pancreative polypeptide. Proc. Natl. Acad. Sci. USA **82:** 1536–1539.

TANABE, T., M. KONISHI, T. MIZUTA, T. NOMA and T. HONJO, 1987 Molecular cloning and structure of the human interleukin 5 gene. J. Biol. Chem. **262:** 16580–16584.

TASANEN, K., T. PARKKONEN, L. T. CHOW, K. I. KIVIRIKKO and T. PIHLAJANIEMI, 1988 Characterization of the human gene for a polypeptide that acts both as the beta subunit of prolyl 4-hydroxylase and as protein disulfide isomerase. J. Biol. Chem. **263:** 16218–16224.

TONE, M., R. KIKUNO, A. KUME-IWAKI and T. HASHIMOTO-GOTOH, 1987 Structure of human alpha-2 plasmin inhibitor deduced from the cDNA sequence. J. Biochem. **102:** 1033–1041.

TONIOLO, D., M. PERSICO and M. ALCALAY, 1988 A "housekeeping" gene on the X chromosome encodes a protein similar to ubiquitin. Proc. Natl. Acad. Sci. USA **85:** 851–855.

TRUONG, A. T., C. DUEZ, A. BELAYEW, A. RENARD, R. PICTET, G. I. BELL and J. A. MARTIAL, 1984 Isolation and characterization of the human prolactin gene. EMBO J. **3:** 429–437.

TSAO, Y.-K., C.-F. WEI, D. L. ROBBERSON, A. M. GOTTO, JR., and L. CHAN, 1985 Isolation and characterization of the human apolipoprotein A-II gene. Electron microscopic analysis of RNA:DNA hybrids, nucleotide sequence, identification of a polymorphic *Msp*I site, and general structural organization of apolipoprotein genes. J. Biol. Chem. **260:** 15222–15231.

TSUJIBO, H., H. F. TIANO and S. S.-L. LI, 1985 Nucleotide sequences of the cDNA and an intronless pseudogene for human lactate dehydrogenase-A isozyme. Eur. J. Biochem. **147:** 9–15.

TSUJINO, S., S. SAKODA, R. MIZUNO, T. KOBAYASHI, T. SUZUKI, S. KISHIMOTO, S. SHANSKE, S. DIMAURO and E. A. SCHON, 1989 Structure of the gene encoding the muscle-specific sub-

unit of human phosphoglycerate mutase. J. Biol. Chem. **264:** 15334–15337.

TSUKADA, T., S. J. HOROVITCH, M. R. MONTMINY, G. MANDEL and R. H. GOODMAN, 1985 Structure of the human vasoactive intestinal polypeptide gene. DNA **4:** 293–300.

UETSUKI, T., A. NAITO, S. NAGATA and Y. KAZIRO, 1989 Isolation and characterization of the human chromosomal gene for polypeptide chain elongation factor-1. J. Biol. Chem. **264:** 5791–5798.

VASICEK, T. J., B. E. MCCEVITT, M. W. FREEMAN, B. J. FENNICK, G. N. HENDY, J. T. POTTS, JR., A. RICH and H. M. KRONENBERG, 1983 Nucleotide sequence of the human parathyroid hormone gene. Proc. Natl. Acad. Sci. USA **80:** 2127–2131.

VERWEIJ, C. L., P. J. DIERGAARDE, M. HART and H. PANNEKOEK, 1986 Full-length von Willebrand factor (vWF) cDNA encodes a highly repetitive protein considerably larger than the mature vWF subunit. EMBO J. **5:** 1839–1847.

WHITE, P. C., M. I. NEW and B. DUPONT, 1986 Structure of human steroid 21-hydroxylase genes. Proc. Natl. Acad. Sci. USA **83:** 5111–5115.

WIGINTON, D. A., D. J. KAPLAN, J. C. STATES, A. L. AKESON, C. M. PERME, I. J. BILYK, A. J. VAUGHN, D. L. LATTIER and J. J. HUTTON, 1986 Complete sequence and structure of the gene for human adenosine deaminase. Biochemistry **25:** 8234–8244.

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

WOOD, W. I., D. J. CAPON, C. C. SIMONSEN, D. L. EATON, J. GITSCHIER, B. KEYT, P. H. SEEBURG, D. H. SMITH, P. HOLLINGSHEAD, K. L. WION, E. DELWART, E. G. D. TUDDENHAM, G. A. VEHAR and R. M. LAWN, 1984 Expression of active human factor VIII from recombinant DNA clones. Nature **312:** 330–337.

XIONG, W., W.-H. LI, I. POSNER, T. YAMAMURA, A. YAMAMOTO, A. M. GOTTO, JR., and L. CHAN, 1991 No severe bottleneck during human evolution: evidence from two apolipoprotein C-II deficiency alleles. Am. J. Hum. Genet. **48:** 383–389.

YOSHIMURA, K., A. TOIBANA and K. NAKAHAMA, 1988 Human lysozyme: sequencing of a cDna, and expression and secretion by *Saccharomyces cerevisiae*. Biochem. Biophys. Res. Commun. **150:** 794–801.

YOSHITAKE, S., B. G. SCHACH, D. C. FOSTER, E. W. DAVIE and K. KURACHI, 1985 Nucleotide sequence of the gene for human factor IX. Biochemistry **24:** 3736–3750.

ZAKUT-HOURI, R., B. BIENZ-TADMOR, D. GIVOL and M. OREN, 1985 Human p53 cellular tumor antigen: cDNA sequence and expression in COS cells. EMBO J. **4:** 1251–1255.

ZIEGLER, S. F., J. D. MARTH, D. B. LEWIS and R. M. PERLMUTTER, 1987 Novel protein-tyrosine kinase gene (HCK) preferentially expressed in cells of hematopoietic origin. Mol. Cell. Biol. **7:** 2276–2285.

Communicating editor: A. G. CLARK