Genetic Evaluation of Suspected Cases of Transient HIV-1 Infection of Infants

Lisa M. Frenkel,* James I. Mullins, Gerald H. Learn, Laura Manns-Arcuino, Belinda L. Herring, Marcia L. Kalish, Richard W. Steketee, Donald M. Thea, Joan E. Nichols, Shan-Lu Liu, Abdallah Harmache, Xi He, David Muthui, Anup Madan, Leroy Hood, Ashley T. Haase, Mary Zupancic, Katherine Staskus, Steven Wolinsky, Paul Krogstad, JiaQi Zhao, Irvin Chen, Richard Koup, David Ho, Bette Korber, Raymond J. Apple, Robert W. Coombs, Savita Pahwa, Norbert J. Roberts Jr.

Detection of human immunodeficiency virus–type 1 (HIV-1) on only one or a few occasions in infants born to infected mothers has been interpreted to indicate that infection may be transient rather than persistent. Forty-two cases of suspected transient HIV-1 viremia among 1562 perinatally exposed seroreverting infants and one mother were reanalyzed. HIV-1 *env* sequences were not found in specimens from 20; in specimens from 6, somatic genetic analysis revealed that specimens were mistakenly attributed to an infant; and in specimens from 17, phylogenetic analysis failed to demonstrate the expected linkage between the infant's and the mother's virus. These findings argue that transient HIV-1 infection, if it exists, will only rarely be satisfactorily documented.

Once an individual is infected with HIV-1, this infection appears to be present for the lifetime of the individual, although the rate of progression to disease may vary. Transient HIV-1 infection has been proposed in a series of cases based on virologic (*1–3*) or immunologic criteria (*4*); however, no cases have been unambiguously confirmed by genetic analyses of the virus (*5*). If transient infection occurred, elucidation of the factors resulting in virus clearance may provide insight into the correlates of protective immunity. The goal of this project was to identify and study individuals transiently infected with HIV-1.

Here a reanalysis is presented of 43 cases of apparent transient viremia (42 infants born to HIV-1–infected mothers and 1 mother). Transient viremia was defined as one or more positive cultures or polymerase chain reaction (PCR) assays for HIV-1 and the subsequent inability to detect HIV-1 in the specimens on multiple occasions or seroreversion, or both. Forty-one cases occurred among 1561 infants in five studies of mother-to-infant HIV-1 transmission. Samples from these 41 cases were obtained from the University of Washington, the North Shore Hospital (*2*), the Pediatric AIDS Clinical Trials Group 076 Study (*6*), the Centers for Disease Control and Prevention's (CDC's) New York Perinatal HIV-1 Transmission Study (*7*), and the Ariel Project (*8*). The infants' specimens taken on the date of the positive virus assay, and their mothers' blood specimens, were analyzed by nested PCR (nPCR) for HIV-1 genes (*9*). Except for specimens from the CDC study, the mothers' and infants' specimens were amplified and sequenced in separate laboratories to exclude the possibility of cross-contamination. Rare somatic genotypes were used to determine whether HIV-1–positive specimens from the CDC study were consistent with the HIV-1–negative specimens from the same infant or if mislabeling of the positive specimen had occurred (*10*). Infants' *env* sequences were compared to their mothers' virus for relatedness by phylogenetic analysis (*11*) (Table 1).

In these five studies of mother-to-infant HIV-1 transmission, infants had been tested a median of three times by culture or PCR for diagnosis of HIV-1 infection. Among these studies, HIV-1–positive specimens occurred in 0.4 to 2.9% of the infants that were perinatally exposed to HIV-1 and were ultimately virus-negative or seroreverting, or both. On reanalysis, viral sequences were not amplified from separate aliquots of previously positive specimens or from other specimens that were available from 16 infants (Table 1). *gag* sequences but no *env* sequences were amplified from four infants' specimens, probably reflecting PCR carryover contamination. Human leukocyte antigen (HLA) typing (*10*, *12*) of infants' specimens revealed mislabeling in five cases, and in these no further viral genetic analyses were done. Viral sequences were detected but no phylogenetic linkage was found between maternal and infant *env* sequences in 15 cases (Figs. 1 and 2) (*12*). These included case 19–208, in which two separate specimens from the same infant were confirmed as virus-positive, neither of which demonstrated close phylogenetic linkage to the mother or to each other (Fig. 1). HIV-1 was amplified from the plasma but not the serum or peripheral blood mononuclear cell (PBMC) DNA of infant 076cE. These plasma sequences formed a monophyletic group with the maternal *env* sequence, differing at the nucleotide level by 7.5% (Fig. 2), which suggested possible epidemiologic linkage between the infant's and mother's viruses (*11*). Somatic genotypes of this infant's plasma, however, were not identical with his sera and PBMCs but shared one haplotype with each maternal locus, which suggested mislabeling of a sibling's specimen.

Specimens were also studied from one mother and her child, both with suspected transient viremia. The mother had two and the infant three positive HIV-1 cultures, but subsequently both individuals became negative for HIV-1 by nPCR, standard virus

L. M. Frenkel, Department of Pediatrics, University of Rochester, Rochester, NY 14642, USA, and Department of Pediatrics, University of Washington, Seattle, WA 98105, USA.
J. I. Mullins, Departments of Microbiology and Medicine, University of Washington, Seattle, WA 98195, USA.
G. H. Learn, B. L. Herring, S.-L. Liu, A. Harmache, X. He, D. Muthui, Department of Microbiology, University of Washington, Seattle, WA 98195, USA.
L. Manns-Arcuino, Department of Pediatrics, University of Washington, Seattle, WA 98195, USA.
R. W. Coombs, Department of Medicine, University of Washington, Seattle, WA 98195, USA.
M. L. Kalish and R. W. Steketee, Centers for Disease Control and Prevention, Atlanta, GA 30333, USA.
D. M. Thea, Medical and Health Research Association of New York City, New York, NY 10013, USA.
J. E. Nichols, Department of Medicine, University of Rochester, Rochester, NY 14642, USA, and Department of Medicine and Microbiology and Immunology, University of Texas Medical Branch, Galveston, TX 77555, USA.
A. Madan and L. Hood, Department of Molecular Biotechnology, University of Washington, Seattle, WA 98195, USA.
A. T. Haase, M. Zupancic, K. Staskus, Department of Microbiology, University of Minnesota, Minneapolis, MN 55455, USA.
S. Wolinsky, Department of Medicine, Northwestern University, Chicago, IL 60611, USA.
P. Krogstad, Department of Pediatrics, University of California at Los Angeles, Los Angeles, CA 90024, USA.
J. Zhao and I. Chen, Department of Medicine and Microbiology and Immunology, University of California at Los Angeles, Los Angeles, CA 90024, USA.
R. Koup and D. Ho, Aaron Diamond AIDS Research Center, New York, NY 10016, USA.
B. Korber, Sante Fe Institute, Sante Fe, NM 87545, USA.
R. J. Apple, Roche Molecular Diagnostics, Alameda, CA 94501, USA.
S. Pahwa, Department of Pediatrics, New York University, North Shore Hospital, Manhasset, NY 11030, USA.
N. J. Roberts Jr., Departments of Pediatrics and Medicine, University of Rochester, Rochester, NY 14642, USA, and Department of Medicine and Microbiology and Immunology, University of Texas Medical Branch, Galveston, TX 77555, USA.

*To whom correspondence should be addressed at the Department of Pediatrics, Division of Infectious Diseases, University of Washington, 4800 Sand Point Way N.E., Box 329500, Seattle, WA 98105, USA. E-mail: lfrenkel@u.washington.edu

cultures, $CD8^+$-depleted virus cultures, and enzyme-linked immunosorbent assay (*12*). HIV-1 RNA and DNA were not detected in two lymph nodes taken from the mother 3 and 4 years after the last virus-positive culture. PCR amplification and DNA sequencing of HIV-1 *env* sequences from the mother's two and the child's three culture supernatants were performed in separate laboratories to eliminate the possibility of cross-contamination. Phylogenetic analysis found that none of the five isolates were genetically linked (Fig. 3) (*12*). Although it is improbable, these five virus isolates appear to have arisen from five separate incidents of specimen contamination or mislabeling. This case remains enigmatic, however, in that both the mother and infant had strong $CD8^+$ cytotoxic T lymphocyte (CTL) responses and lymphocyte proliferation to multiple HIV-1 antigens (*12*).

In all 43 cases of suspected transient HIV-1 infection reported here, transient viremia was ruled out by genetic criteria. PCR failed to confirm the virus in specimens from 20 cases, and these were considered to be falsely positive in the initial tests. Typing of somatic loci demonstrated nonconcordance between genotypes from the virus-positive and virus-negative specimens attributed to six infants (five infants from the CDC study and infant 076cE), which indicates that specimen mislabeling was responsible for the positive test result. Phylogenetic analysis of *env* sequences failed to identify the expected monophyletic relationship or linkage between infant and maternal viral sequences in 17 cases, which indicates specimen mislabeling or laboratory contamination. Furthermore, resistance to HIV-1 infection on the basis of a homozygous deletion within the *CCR5* gene (*13*) was also excluded in all 29 infants tested. Thus, each case of suspected viremia we investigated appeared to be due to sample mislabeling or laboratory contamination.

Previous reports proposing transient HIV-1 infection have presented either virologic or immunologic evidence of infection. In one case (*1*) HIV-1 sequences from two of an infant's specimens taken at different times were phylogenetically linked with each other; however, linkage was not established to virus from the infant's mother (*5*, *14*). Although rare somatic genetic markers linked the child to stored PBMCs taken at the same time as one of his HIV-1–positive cultures, the presence and genetic characterization of viral sequences from the PBMCs were not reported. Thus, the lack of phylogenetic linkage between viral sequences ascribed to the mother and child, and the lack of confirmation of virus in banked PBMCs, serum, or plasma specimens genetically linked to the child, demonstrate the difficulty of proving transient infection.

Less well studied were 15 infants reported by two different groups (*2*, *3*). HIV-1 DNA was amplified from the PBMCs from each child on two to seven occasions during the first year of life, but the infants were all ultimately determined to be uninfected with HIV-1. We evaluated specimens from two of these cases (*2*) and were unable to confirm

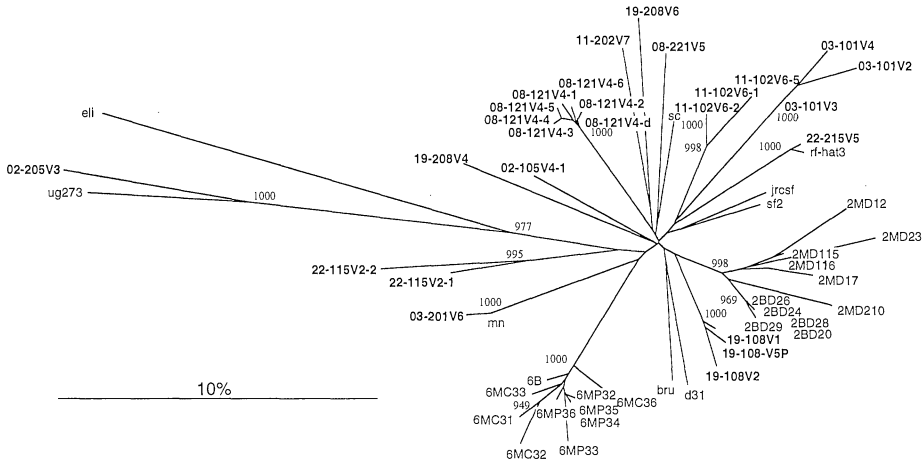

Fig. 1. Neighbor joining phylogenetic tree of HIV-1 DNA sequences from cases of suspected transient HIV-1 infection from the Ariel Project cohort. Nomenclature for each child sequence is as indicated in Table 1 (the XX-2XX series), with the corresponding mother sequence using a similar identifier (the XX-1XX series). Examples taken from another study of monophyletic linked specimens from two confirmed cases of mother-to-infant HIV-1 transmission are shown (the 6B/6MXXX series and the 2BXXX/2MXXX series) (*23*). In Figs. 1 through 3, the scale bar indicates branch lengths corresponding to 10% sequence divergence. Additional sequences shown in Figs. 1 through 3 are taken from GenBank. These GenBank sequences correspond to common laboratory strains, plus some sequences that illustrate the lack of monophyly between mother and infant pairs—sequences that are more closely related to a mother's virus than are those attributed to her infant: bru, K02013; cam, D10112; d31, U43096; eli, K03454; jrcsf, M38429; jrfl, U63632; ma6, M79352; mn, M17449; ny5, M38431; oyi, M26727; rf-hat3, M17451; sf2, K02007; sc, M17450; th14, U08801; tp3, U95452; ug273, L22957; yu10, M93259.

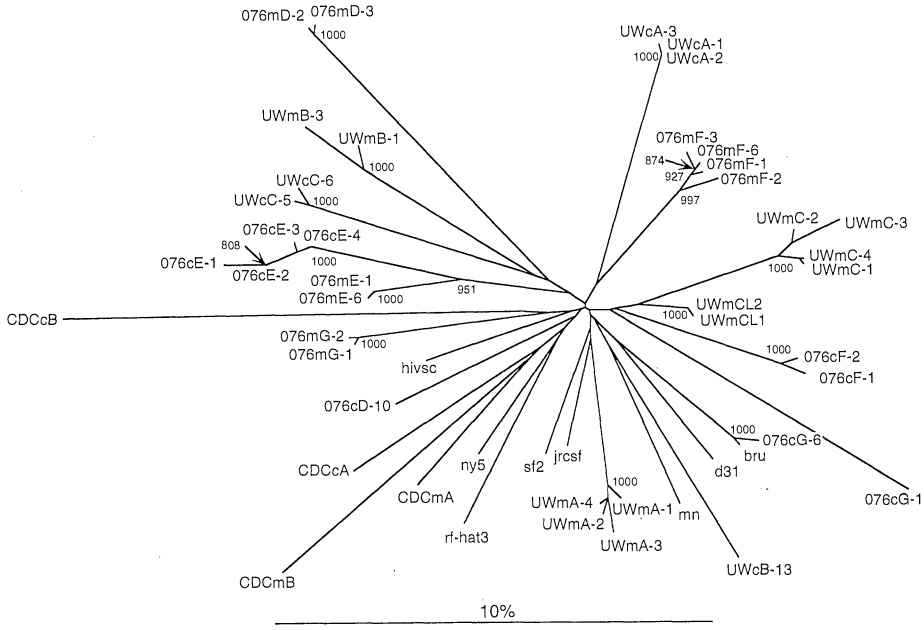

Fig. 2. Neighbor joining phylogenetic tree of HIV-1 DNA sequences from cases of suspected transient HIV infection from the UW, ACTG 076, and CDC New York cohorts. The nomenclature for each child sequence is as indicated in Table 1, and maternal virus sequences are indicated with the same nomenclature, substituting "m" for "c" in the name. No sequence is shown from mother-child pair 076cH (see Table 1) because of a shorter read length.

Table 1. Laboratory test results that suggested transient HIV-1 viremia and subsequent testing of infant specimens are shown. Reanalysis did not support transient infection: HIV-1 *env* could not be amplified from 20 cases, and somatic genotyping and phylogenetic analysis revealed specimen mislabeling or contamination in the remaining 23 cases. UW, the University of Washington; North Shore, the North Shore Hospital, ACTG 076, the Pediatric AIDS Clinical Trials Group 076 Study.

Study name and participant number	Initial tests: Age (weeks) at positive test			PCR reanalysis of specimens from date of positive test							Somatic genotype	Phylogenetic linkage
				Plasma			PBMC5		Serum	Culture		
	DNA PCR	RNA PCR	PBMC culture	*env* C2-V5	*gag*	*env* gp160	*env* C2-V5	*gag*	*env* C2-V5	*env* C2-V5		
Mother and child												
Mother (3'91)										+		−
Mother (6'91)										+		−
Child (3'90)										+		−*
Child (5'90)										+		−
Child (9'90)				−	−					+		−
UW cohort												
UWcC			25							+		−
UWcB			16							+		−
UWcA			18	−			−			+		−
UWc1		36			−							
North Shore												
NSc151	0.3			−	+/−†							
	16			−	−							
NSc288	4			−	−							
	11						−	−				
	11.13			−	−							
ACTG 076												
076cH			12	−			−		+			−
076cG			0	−			+		−	+		−
076c2			24	−								
076cF			0	−			−		+			−
076cE			24	+			−		−			−
076cD			24	−			+		−			−
CDC NY study												
CDC66922 (cA)	14				−‡		+§				+	−
CDC66937	15										−	
CDC91741	1				−‖		−¶				+	
CDC804199	0.13										−	
CDC804267 (cB)	2						+‖				+	−
CDC800210	14				+						−	
CDC804270	4										−	
CDC892921	0.13										−	
Ariel Project												
02-202	18					−	−	+				
02-204	0.13					−		−				
02-205	6					−	+	+				−
02-206	0.13					−		−				
02-232A	26							−				
02-237	6							−				
	10#							−				
02-238	0.13							−				
	1				−	−	−	+				
02-240	26							−				
03-201	10		10		−	+						−**
08-206	38					−		−				
08-218	26					−		−				
08-221			18		−		+					−
11-202	0.13#					−		−				
			38		−		+	+				−
11-208	10				−	−		−				
	18#					−		−				
	26				−							
	54#				−							
18-202	54					−		−				
18-205	6					−		−				
19-206	18				−	−		−				
19-208			10		−		+					−
	26				−		+	+				−
22-214	10#					−	−	+				
22-215	18					−	+	+				−††
22-239	0.13#					−		−				
	6#				−	−		−				
	26				−							

*Sequence virtually identical to the laboratory strain HIV-1-Lai/IIIB. †Two of 10 PCR reactions positive; sequence found to be *gag* from clade A. ‡The NASBA QT assay kit was used for detection. §The NASBA QL assay kit was used for detection. ‖PCR was done over V3 plus flanking regions (*21*). ¶PCR was done with the use of several sets of *env* PCR primers (*22*). #Indeterminate PCR, indicating a weakly positive signal. ** This PBMC sample was initially PCR-negative but was positive upon repeat testing. Phylogenetic analysis showed that it was not linked to the mother's virus but was virtually identical to the laboratory strain HIV-1-MN. ††Virtually identical to the laboratory strain HIV-1-RF.

the presence of HIV-1 nucleic acids.

Transient HIV-1 infection of infants has also been argued on the basis of multiple positive tests for HIV-1, which would be predicted to occur rarely (*15*). This analytic approach assumed the independence of each test result and thus did not recognize the nature of PCR carry-over contamination, which may cluster within laboratories. Also, the amplification of only one of three HIV-1 genes was used to define a positive PCR test; this favors the categorization of carry-over contamination as infection. Our studies demonstrate that claims of transient infection based only on multiple positive tests for HIV-1 in an infant are not valid.

It has also been suggested that transient infection occurred in the wife of an HIV-1–infected individual, based on the fact that HIV-1 *gag* was found in PBMC DNA on one occasion (*16*). On reanalysis, this specimen was linked to the woman by somatic genetic markers, *gag* was reamplified, and *env* was amplified. However, HIV-1 RNA was not detected in the specimen, and testing of a separate aliquot of the specimen was not reported. Furthermore, the phylogenetic relationship of the woman's virus to the purported source of her infection, her husband, was not evaluated.

We suggest that substantiation of transient infection requires demonstration of genetically linked viral strains in the donor and the recipient by phylogenetic analysis of viral DNA (*11*). Studies of the donor and recipient specimens should use previously unmanipulated aliquots of the specimens, the identity of which should be proven by somatic markers (*10*). Ideally, separate laboratories should evaluate the donor's and recipient's specimens with reagents prepared separately by different personnel. We disagree with the recommendation that heteroduplex techniques be used (*17*), which would involve concomitant handling of PCR products from the donor and recipient. Linkage between presumed donor and recipient specimens should be verified by analyses of more than one separate phylogenetically informative region of the viral genome, such as *env* plus the *gag* p17 coding sequence (*18*), preferably using a separate aliquot of the specimen, to help rule out PCR carry-over contamination.

The utility of immunologic testing as evidence for past transient HIV-1 infection requires further substantiation. For example, the significance of specific cell-mediated immune responses detected in individuals who have had known exposures to HIV-1 but have remained seronegative and HIV-1–negative by culture or PCR analysis (*4*) should be critically investigated. In addition, such studies should involve concurrent testing of matched HIV-1–unexposed control specimens. HIV-1–specific CTLs are generally thought to result from antigens synthesized within the presenting cell and thus to require viral infection. However, several observations suggest that the presence of specific CTLs may not require a productive infection. For example, exposure to nonreplicating Sendai virus has been observed to result in specific CTLs (*19*). A strong cellular immune response was detected in both the mother and child of the pair we reported who had suspected transient viremia, and although the mother reported sexual exposures to HIV-1–infected partners, the child had no known exposure to HIV-1 (*12*). HIV-1–specific CTL precursor cells also have been detected in other people not known to have been exposed to HIV-1 (*20*). The events eliciting HIV-1–specific CTLs and other cellular immune reactivity among these uninfected and presumably unexposed individuals require definition.

Unlike immunologic studies, genetic analyses can directly and critically evaluate cases of suspected transient viremia. No case thus far reported has satisfied the viral and host genetic characterization criteria described here for the verification of transient HIV-1 infection. Our negative studies of 43 cases of suspected transient infection indicate that the phenomenon of transient HIV-1 infection remains to be proven and that most cases suggestive of transient HIV-1 infection are cases of mislabeling of specimens or their contamination in the laboratory.

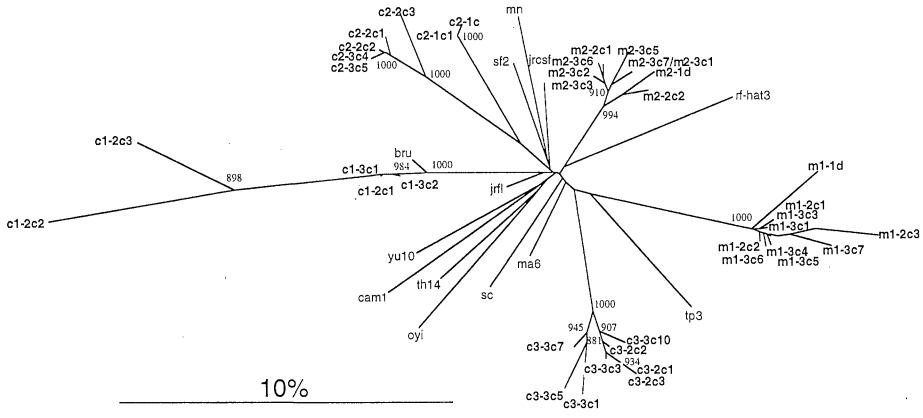

Fig. 3. Neighbor joining phylogenetic tree of HIV-1 DNA sequences from the mother-child pair in which both had apparent transient viremia. Five viral isolates are shown: three from the infant and two from the mother. No phylogenetic linkage was observed between these five isolates, although clones generated from three aliquots of each isolate demonstrated linkage within the isolate. A maximum likelihood phylogenetic analysis of representative sequences from the broader analysis shown here allowed us to reject the hypothesis of monophyly for the viral sequences attributed to the mother and infant (*14*) (P = 0.0003, Kishino-Hasegawa test). A detailed description of this case, including the virologic and immunologic analysis, can be found at *Science* Online (*12*). Each sequence is identified by "c" for child or "m" for mother. m1 is from a sample attributed to the mother in March 1991; m2, in June 1991. c1 is from a sample attributed to the child in March 1990; c2, in May 1990; c3, in September 1990. The aliquot number and clone number (XcX) follow the hyphen. Bootstrap values (*11*) >800/1000 are reported at the nodes.

REFERENCES AND NOTES

1. Y. J. Bryson *et al.*, *N. Engl. J. Med.* **332**, 833 (1995).
2. S. S. Bakshi, S. Tetali, E. J. Abrams, M. O. Paul, S. G. Pahwa, *Pediatr. Infect. Dis. J.* **14**, 658 (1995).
3. P. A. Roques *et al.*, *AIDS 1995* **9**, F19 (1995).
4. W. Borkowsky, K. Krasinski, T. Moore, V. Papaevangelou, *AIDS Res. Hum. Retroviruses* **6**, 673 (1990); M. Clerici *et al.*, *J. Infect. Dis.* **165**, 1012 (1992); R. Cheynier *et al.*, *Eur. J. Immunol.* **22**, 2211 (1992); M. Clerici *et al.*, *AIDS* **7**, 1427 (1993); S. L. Rowland-Jones *et al.*, *Lancet* **341**, 860 (1993); S. L. Rowland-Jones *et al.*, *Nature Med.* **1**, 59 (1995); M. Clerici *et al.*, *JAMA* **271**, 42 (1994); A. De Maria, C. Cirillo, L. Moretta, *J. Infect. Dis.* **170**, 1296 (1994).
5. M. O. McClure *et al.*, *Nature* **375**, 637 (1995).
6. E. M. Connor and L. M. Mofenson, *Pediatr. Infect. Dis. J.* **14**, 536 (1995).
7. R. W. Steketee *et al.*, *J. Infect. Dis.* **175**, 707 (1997).
8. Y. Cao *et al.*, *Nature Med.* **3**, 549 (1997).
9. L. M. Frenkel *et al.*, *Clin. Infect. Dis.* **20**, 1321 (1995). Reaction conditions plus a detailed list of PCR primers covering the *gag*, *pol*, and *env* regions used for assessing HIV-1 infection are provided in (*12*). For phylogenetic analysis, a 650–base pair (bp) region of HIV-1 *env*, corresponding to the V3 through V5 region, was amplified by nPCR with primers ED5/12 and ES7/8 [E. L. Delwart *et al.*, *Science* **262**, 1257 (1993)] and cloned into plasmids. Four to six clones from each specimen were sequenced. In the case of the CDC specimens, direct sequencing of 345 bp of the V3 and flanking regions was done as previously described [H. W. Jaffe *et al.*, *Ann. Intern. Med.* **121**, 855 (1994)].
10. Independently segregating loci or the HLA-DQa locus were used to evaluate the identity of patient specimens. The former included HLA-DQA1, low-density lipoprotein receptor, glycophorin A, hemoglobin G gammaglobin, D7S8, and a group-specific component (GC) (AmpliType PM PCR Amplification and Typing Kit, Perkin-Elmer, Foster City, CA). The combined power of discrimination for these six markers for Caucasians in the United States is 0.9997. Discordance of the HLA-DQA1 loci (AmpliType HLA DQa Amplification and Typing Kit, Perkin-Elmer) among the CDC infant's specimens was taken to indicate specimen mislabeling. When an infant's specimens had concordant HLA-DQa loci, the

env sequences in each specimen were examined by phylogenetic analysis.
11. The methods described by G. H. Learn *et al.* [*J. Virol.* **70**, 5720 (1996)] were used to align DNA sequences (with the use of CLUSTALW plus manual adjustment), calculate genetic distances (with the use of DNADIST, using the maximum likelihood method), evaluate potential sample mixups, construct neighbor joining trees, and perform bootstrap analyses (1000 replicates). Sequence regions that could not be unambiguously aligned were removed from subsequent analyses. Each sequence was compared for phylogenetic relatedness to the entire set of published and available unpublished laboratory HIV database sequences. If after this analysis the viral sequences from a mother and an infant appeared as a monophyletic group on a phylogenetic tree, they were judged to be phylogenetically linked or to have a common ancestor not shared by sequences from any other individuals evaluated. Issues regarding the assignment of phylogenetic linkage are discussed in greater detail by Learn *et al.*
12. L. M. Frenkel *et al.*, at www.sciencemag.org/feature/data/974996.shl.
13. R. Liu *et al.*, *Cell* **86**, 367 (1996).
14. L. M. Frenkel *et al.*, unpublished data.
15. M.-L. Newell *et al.*, *Lancet* **347**, 213 (1996).
16. P. Palumbo, J. Skurnick, D. Lewis, M. Eisenberg, *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.* **10**, 436 (1995).
17. A. McMichael, R. Koup, A. J. Ammann, *N. Eng. J. Med.* **334**, 801 (1996).
18. E. C. Holmes *et al.*, *J. Infect. Dis.* **167**, 1411 (1993).
19. T. Liu *et al.*, *J. Immunol.* **154**, 3147 (1995).
20. A. Hoffenbach *et al.*, *ibid.* **142**, 452 (1989).
21. G. Schochetman, S. Subbarao, M. L. Kalish, in *Viral Genome Methods*, K. W. Adolph, Ed. (CRC Press, Boca Raton, FL, 1996), pp. 25–41.
22. E. L. Delwart, M. P. Busch, M. L. Kalish, J. W. Mosley, J. I. Mullins, *AIDS Res. Hum. Retrovir.* **11**, 1181 (1995).
23. C. H. Contag *et al.*, *J. Virol.* **71**, 1292 (1997).
24. We thank J. Conroy for performing PCR assays; E. Abrams, M. S. Orloff, R. C. Reichman, L. M. Demeter, J. S. Lambert, R. Dolin, R. Sperling, D. Shapiro, G. McSherry, and the Ariel Project and ACTG 076 investigators for critical patient specimens; D. Swofford for use of computer program PAUP*, version 4.0.0d63; and C. B. Wilson and K. K. Holmes for editorial contributions. This work was supported by grants from the Pediatric AIDS Foundation (500153–10-PGT, 50366–14-PGR, 55516-ARI, 55529-ARI, 55525-ARI, 55532-ARI, 55526-ARI, 55531-ARI, and 55522-ARI), the U.S. Public Health Service (UO1–27658, AI32910, AI27757, and AI35539), and the Foster Foundation.

22 December 1997; accepted 26 March 1998

Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome

David G. Wang, Jian-Bing Fan, Chia-Jen Siao, Anthony Berno, Peter Young, Ron Sapolsky, Ghassan Ghandour, Nancy Perkins, Ellen Winchester, Jessica Spencer, Leonid Kruglyak, Lincoln Stein, Linda Hsie, Thodoros Topaloglou, Earl Hubbell, Elizabeth Robinson, Michael Mittmann, Macdonald S. Morris, Naiping Shen, Dan Kilburn, John Rioux, Chad Nusbaum, Steve Rozen, Thomas J. Hudson, Robert Lipshutz,* Mark Chee, Eric S. Lander*

Single-nucleotide polymorphisms (SNPs) are the most frequent type of variation in the human genome, and they provide powerful tools for a variety of medical genetic studies. In a large-scale survey for SNPs, 2.3 megabases of human genomic DNA was examined by a combination of gel-based sequencing and high-density variation-detection DNA chips. A total of 3241 candidate SNPs were identified. A genetic map was constructed showing the location of 2227 of these SNPs. Prototype genotyping chips were developed that allow simultaneous genotyping of 500 SNPs. The results provide a characterization of human diversity at the nucleotide level and demonstrate the feasibility of large-scale identification of human SNPs.

Although the Human Genome Project still has tremendous work ahead to produce the first complete reference sequence of the human chromosomes, attention is already focusing on the challenge of large-scale characterization of the sequence variation among individuals (*1*). This genetic diversity is of interest because it explains the basis of heritable variation in disease susceptibility, as well as harbors a record of human migrations.

The most common type of human genetic variation is the SNP, a position at which two alternative bases occur at appreciable frequency (>1%) in the human population. There has been growing recognition that large collections of mapped SNPs would provide a powerful tool for human genetic studies (*1*, *2*). SNPs can serve as genetic markers for identifying disease genes by linkage studies in families, linkage disequilibrium in isolated populations, association analysis of patients and controls, and loss-of-heterozygosity studies in tumors (*1*, *2*). Although individual SNPs are less informative than currently used genetic markers (*3*), they are more abundant and have greater potential for automation (*4*, *5*).

We performed an initial survey to identify SNPs by using conventional gel-based DNA sequencing to examine sequence-tagged sites (STSs) distributed across the human genome. STSs are short genomic sequences that can be amplified from DNA samples by means of a corresponding polymerase chain reaction (PCR) assay. From among 24,568 STSs used in the construction of a physical map of the human genome at the Whitehead Institute for Biomedical Research/MIT Center for Genome Research (*6*, *7*), an initial collection of 1139 STSs was chosen (*8*). These STSs contained a total of 279 kb of genomic sequence (*9*), with one-third from random genomic sequence and two-thirds from 3′-ends of expressed sequence tags (3′-ESTs) and primarily representing untranslated regions of genes. Each STS was amplified from four samples (*10*): three individual samples and a pool of 10 individuals (thereby permitting allele frequencies to be estimated among 20 chromosomes). The PCR products were subjected to single-pass DNA sequencing based on fluorescent-dye primers and gel electrophoresis; sequence traces were compared by a computer program followed by visual inspection (*11*). Candidate SNPs were declared when two alleles were seen among the three individuals, with both alleles present at a frequency greater than 30% in the pooled sample. The term "candidate SNP" is used because a subset of such apparent polymorphisms turn out to be sequencing artifacts, as discussed below.

The survey identified 279 candidate SNPs, distributed across 239 of the STSs. This corresponds to a rate of one SNP per 1001 base pairs (bp) screened and an observed nucleotide heterozygosity of $H = 3.96 \times 10^{-4}$ (Table 1). Expressed sequences (3′-ESTs) showed a lower polymorphism rate than random genomic sequence (with

D. G. Wang, C.-J. Siao, P. Young, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, E. Robinson, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T. J. Hudson, Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142, USA.
J.-B. Fan, A. Berno, R. Sapolsky, G. Ghandour, L. Hsie, T. Topaloglou, E. Hubbell, M. Mittmann, M. S. Morris, N. Shen, R. Lipshutz, M. Chee, Affymetrix, Incorporated, 3380 Central Expressway, Santa Clara, CA 95051, USA.
E. S. Lander, Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142, USA, and Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
*To whom correspondence should be addressed.