Am. J. Hum. Genet. 66:674–686, 2000

Y Chromosomes Traveling South: The Cohen Modal Haplotype and the Origins of the Lemba—the "Black Jews of Southern Africa"

Mark G. Thomas,[1] Tudor Parfitt,[3] Deborah A. Weiss,[4] Karl Skorecki,[5] James F. Wilson,[2] Magdel le Roux,[6] Neil Bradman,[7] and David B. Goldstein[2]

[1]The Center for Genetic Anthropology, Departments of Biology and Anthropology, and [2]Galton Laboratory, Department of Biology, University College London, and [3]School of Oriental and African Studies, University of London, London; [4]Department of Anthropology, University of California, Davis; [5]Bruce Rappaport Faculty of Medicine and Research Institute, Technion and Rambam Medical Center, Haifa, Israel; [6]Department of Old Testament, University of South Africa, Pretoria; and [7]Department of Zoology, University of Oxford, Oxford

Summary

The Lemba are a traditionally endogamous group speaking a variety of Bantu languages who live in a number of locations in southern Africa. They claim descent from Jews who came to Africa from "Sena." "Sena" is variously identified by them as Sanaa in Yemen, Judea, Egypt, or Ethiopia. A previous study using Y-chromosome markers suggested both a Bantu and a Semitic contribution to the Lemba gene pool, a suggestion that is not inconsistent with Lemba oral tradition. To provide a more detailed picture of the Lemba paternal genetic heritage, we analyzed 399 Y chromosomes for six microsatellites and six biallelic markers in six populations (Lemba, Bantu, Yemeni-Hadramaut, Yemeni-Sena, Sephardic Jews, and Ashkenazic Jews). The high resolution afforded by the markers shows that Lemba Y chromosomes are clearly divided into Semitic and Bantu clades. Interestingly, one of the Lemba clans carries, at a very high frequency, a particular Y-chromosome type termed the "Cohen modal haplotype," which is known to be characteristic of the paternally inherited Jewish priesthood and is thought, more generally, to be a potential signature haplotype of Judaic origin. The Bantu Y-chromosome samples are predominantly (>80%) YAP$^+$ and include a modal haplotype at high frequency. Assuming a rapid expansion of the eastern Bantu, we used variation in microsatellite alleles in YAP$^+$ sY81-G Bantu Y chromosomes to calculate a rough date, 3,000–5,000 years before the present, for the start of their expansion.

Received May 25, 1999; accepted November 23, 1999; electronically published February 11, 2000.

Address for correspondence and reprints: Prof. David B. Goldstein, Galton Laboratory, Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE. E-mail: d.goldstein@ucl.ac.uk

0002-9297/2000/6602-0034$02.00

Introduction

The Lemba, once referred to as "Kruger's Jews" (because President Paul Kruger, President of Transvaal during 1883–1900, was thought to have discovered them), are commonly referred to as the "black Jews" of South Africa. Their claim of Jewish origin is based on slim evidence: a persistent oral tradition of uncertain antiquity and a number of suggestive customs, from circumcision to food taboos, which appear to be "Judaic" but could be Muslim or, indeed, in the case of circumcision, African (Mandivenga 1983). Lemba tradition holds that the tribe came from "Sena in the north by boat." The original group is said to have been entirely male, with half of their number having been lost at sea; the remainder made their way to the coasts of Africa. Once there, they rebuilt their city of Sena, later leaving it to build a second city of the same name. "Sena" is variously identified by the Lemba as Sanaa in Yemen, Judea, Egypt, or Ethiopia (Ruwitah 1997; Parfitt 1997). The first clear and unambiguous reference to the Lemba as a separate tribe and perhaps polity is from a Dutch report from 1721 (Liesenbang 1977). Today the religious life of the Lemba is highly syncretistic. Many of them belong to various Christian churches (e.g., the Zion Christian Church and Pentecostal groups), whereas some in Zimbabwe are Muslims. Others, however, claim to be Lemba by religious practice as well as by ethnic identification. The religious practices of these Lemba do not have much in common with Judaism as it is practiced elsewhere. There are thought to be ~50,000 Lemba spread over South Africa and Zimbabwe, with some closely connected groups in Malawi (Parfitt 1997). At some time in the past they became scattered among the more powerful neighboring tribes, where they served particularly as "medicine men," iron and copper workers, traders, and officials with ritual responsibilities. They traded throughout southern Africa.

The Lemba have >12 clans, some of which appear to correlate with place names in the Hadramaut (Parfitt 1997). The Buba clan is recognized as being the senior

clan, both the oldest and, for some ritual purposes, the most important. Parfitt (1997) has claimed to have discovered the original Sena of the Lemba in the eastern Hadramaut in the Yemen.

Y-Chromosome Markers

The paternally inherited nonrecombining portion of the Y chromosome includes polymorphisms that mutate relatively frequently (microsatellites) and biallelic polymorphisms (the YAP *Alu* insert and single-nucleotide substitutions) that are unlikely to have arisen more than once in human evolution. For this reason, the latter group has been called "unique event polymorphisms" (UEP [Thomas et al. 1998]), and it can usually be assumed to partition Y chromosomes into distinct genealogical groups. The ability to identify high-resolution haplotypes comprising linked markers mutating at different rates makes the Y chromosome a powerful instrument for investigation of relationships between geographically distant populations that may have been obscured through extensive admixture with their neighbors (Thomas et al. 1998).

A previous study compared the frequency of RFLP polymorphisms p12F2, p49a, and pDP31 and the YAP insert in the Lemba and a number of other populations (Spurdle and Jenkins 1996). The distribution of haplotypes identified was considered to be consistent with Lemba oral tradition but could not distinguish a Jewish from a more general Semitic contribution to the Lemba gene pool. Markers used in this study comprise six microsatellites (DYS19, DYS388, DYS390, DYS391, DYS392, and DYS393) and six UEP (YAP, SRY4064, sY81, SRY+465, 92R7, and Tat). The 12 polymorphisms listed above were characterized in multiple Jewish populations and identified a single haplotype (Cohen modal haplotype [CMH]) that is dominant (frequency ~.50) in the Jewish priesthood and that may be more generally characteristic of Hebraic ancestry. Although the frequency is moderate (~.12) in lay Jews, the CMH is absent or at low frequency in Yakut, Mongolians, Nepalese, Armenians, Greeks, and Cypriots (authors' unpublished data) and, interestingly, in Palestinian Arabs (A. Nebel, D. Filon, M. Faerman, A. Oppenheim, personal communication). The combination of the presence of the CMH at high frequency in the Lemba and its absence in neighboring Bantu populations would be supportive of Lemba claims of a paternal Judaic ancestry, especially if its frequency is relatively low in other Semitic groups.

Subjects And Methods

Study Populations

Lemba.—Samples were collected from paternally unrelated but otherwise random males, identified as Lemba by the subjects themselves, on two separate occasions, in the Louis Trichardt area of Northern Province (January 1997; 90 samples) and in Sekhukuneland in Mpumalanga (October 1997; 46 samples), South Africa. Clan affiliations of 108 (.794) of the subjects were recorded.

Bantu.—Samples were collected (February 1998) from paternally unrelated males of various chieftainships in the Pretoria area of South Africa who spoke a Bantu language. The subjects included two men born in Senegal who spoke Wolof (a non-Bantu language) as their first tongue. There is concern over the use of the word "Bantu" (Cavalli-Sforza et al. 1994, p. 185). We follow Cavalli Sforza et al. in noting that, although "Bantu" was originally a linguistic term, its use to define population groups can be justified on the assumption that a geographic expansion spread both the Bantu language and a group of related people.

Yemeni.—Collections were made (May 1997) at two locations in the Hadramaut, Yemen, from paternally unrelated but otherwise randomly selected males. Specific locations were the Seiyun Teachers Training College in Seiyun in the Hadramaut and Sena, a small (population ~3,000) isolated town located ~60 km east of Terim and ~40 km from the coast. The college draws its membership from a local but dispersed area and forms part of the new University of the Hadramaut. The present-day residents of Sena suggest that their ancestors may have moved to the area relatively recently (<300 years ago).

The Hadramaut population is known to have a prolonged history of seafaring and trading with eastern Africa and other overseas areas, including Southeast Asia, and has a considerable diaspora in Indonesia, eastern Africa, Saudi Arabia, and Egypt. The people are Arab Muslims, although Jews are thought to have lived in the region during former times (Parfitt 1997).

Ashkenazic Israelites and Sephardic Israelites.—Samples were collected from self-designated, paternally unrelated but otherwise randomly selected males in Canada, the United States, the United Kingdom, and Israel. In all cases, appropriate informed consent was obtained before samples were collected.

Today, Jewish males can be divided into three castes: Cohanim (the paternally inherited priesthood), Leviim (non-Cohen members of the paternally defined priestly tribe of Levi), and Israelites (all non-Cohen and non-Levite Jews). Significant differences in Y-chromosome frequencies among these groups were recently reported by Thomas et al. (1998). As a consequence, frequencies of Y-chromosome haplotypes of Jewish populations are expected to vary, in part because of differences in the proportion of Cohanim, Leviim, and Israelites sampled. Since ~90% of the Jewish population are Israelites (Bradman et al., in press), it was decided, for the purpose of the current study, that, in terms of the data already available, Jewish populations are best represented by

Table 1

Distribution of Y Chromosomes from Six Populations into Four UEP Groups, and Nei's Genetic Identity

	Frequency (*n*) in					
	AI	SI	Y	S	L	B
UEP group:[a]						
1 YAP⁻GACCT	.650 (39)	.620 (31)	.735 (36)	1.000 (27)	.654 (89)	.169 (13)
2 YAP⁻GACTT	.183 (11)	.240 (12)	.163 (8)	.000 (0)	.015 (2)	.000 (0)
3 YAP⁺AACCT	.167 (10)	.140 (7)	.061 (3)	.000 (0)	.029 (4)	.026 (2)
4 YAP⁺AGCCT	.000 (0)	.000 (0)	.041 (2)	.000 (0)	.302 (41)	.805 (62)
	1.000 (60)	1.000 (50)	1.000 (49)	1.000 (27)	1.000 (136)	1.000 (77)
Nei's genetic identity I:						
SI	.995					
Y	.984	.980				
S	.934	.913	.972			
L	.863	.844	.912	.907		
B	.199	.194	.255	.205	.596	

NOTE.—AI = Ashkenazic Israelites, SI = Sephardic Israelites, Y = Yemeni, S= Sena, L = Lemba, B = Bantu, A = adenine, C = cytocine, G = guanine, and T = thymine.

[a] Polymorphisms in the order YAP, SRY4064, sY81, SRY+465, 92R7, Tat.

males of the Israelite caste. Data on some of the Ashkenazic- and Sephardic-Israelite samples have been the subject of publication elsewhere (Skorecki et al. 1997; Thomas et al. 1998).

Population Designations

Ashkenazic.—These Jews follow the Ashkenazic rite associated with the Jewish communities of northern Europe.

Sephardic.—These Jews follow a Sephardic rite associated with the Jewish communities of northern Africa and Asia. The two communities have, to a great extent, been isolated from each other for ⩾500 years.

Semitic.—Unless otherwise noted, "Semitic" here will refer to the Jewish, Yemeni, and Sena populations together.

Yemeni and Sena.—Samples collected at the Seiyun Teachers Training College are referred to here as "Yemeni," samples collected at Sena as "Sena."

Sample Collection and Extraction of DNA

In sample collection from Ashkenazic and Sephardic Jews, the subject lightly scraped the inside of the cheek with a wooden spatula, swished saline solution around the mouth for ~30 s, and expelled the solution into a collection tube. Sample collection from all other groups was by buccal swabs wiped over the inside of the cheek and returned to collection tubes to which 0.05 M EDTA/0.5% SDS was added. DNA was extracted by use of a standard phenol procedure.

Typing

Two multiplex kits (one for the six microsatellites and the other for the six UEP) were developed for use with an ABI-310 genetic analyzer (Thomas et al. 1998). Five percent of the samples were retyped for microsatellite repeats on an ABI-377 genetic analyzer in a separate laboratory. Only 1/120 typings were different (DYS392; a difference of two "steps").

Genetic Distances, Haplotypes, and Genealogical Trees

Genetic distances for haplotypes were calculated with use of the computer program MICROSAT, to infer genealogical trees that were drawn with the program PHYLIP (Felsenstein 1993). Genetic similarities of populations were calculated by use of Nei's (1987) genetic identity. Microsatellite haplotypes present at a frequency ⩾.1 in any of the six populations were included in a table of frequently encountered haplotypes.

Results

UEP Groups

Table 1 classifies the Y chromosomes of the six sample populations into four groups (for a cladogram illustrating the evolutionary relationships of the four UEP groups, see fig. 1). The distribution of the four UEP groups among the populations is apparently consistent with earlier suggestions that the Lemba have a mixed Bantu-Semitic origin. The frequency of the YAP⁺ chromosomes in the Lemba is intermediate between that of the Bantu group and that of the Semitic group. However, the distribution between the two YAP⁺ groups is significantly different. Only two Bantu, four Lemba, and three Yemeni Y chromosomes are included in UEP group 3 (YAP⁺ sY81-A), whereas all of the YAP⁺ Israelites fall into this group. YAP⁺ sY81-A Y chromosomes have been identified in many widely separated populations in Asia, Africa, and Europe. It is more widespread than the derived YAP⁺sY81-G haplotype (UEP group 4), which ap-

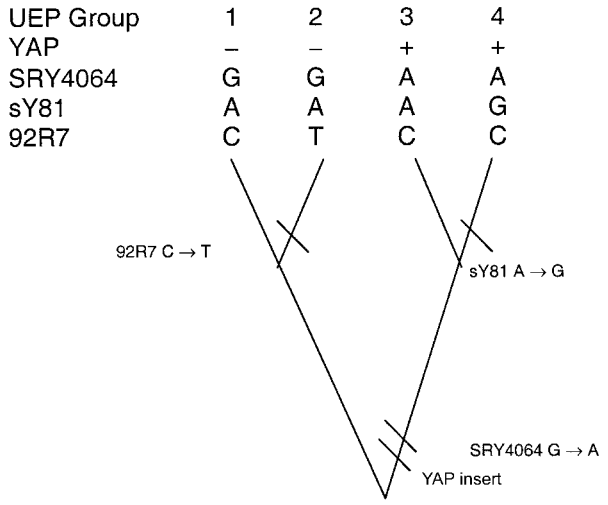

Figure 1 Cladogram of UEP, distinguishing UEP groups. Explanations of abbreviations are as in table 1.

pears to be substantially confined to Africa (Hammer et al. 1997). Given the significant difference between the distribution of the two YAP^+ UEP groups in the Semitic and Bantu populations, interpreting the total frequency of YAP^+ Y chromosomes as an indicator of admixture in the Lemba could be misleading. Distinguishing the frequencies of the YAP^+ sY81-A and YAP^+ sY81-G Y chromosomes, however, seems to provide a more appropriate level of resolution.

All members of the Sena population fall within UEP group 1, the most common UEP group (YAP^- 92R7-C) in the Semitic populations. The relatively greater homogeneity of the Sena population may result from the town's isolation and from its catchment area being more limited than those for the other Hadramaut samples. Although the Ashkenazic Israelites, Sephardic Israelites, and Yemeni are not significantly different from each other in the distribution of Y chromosomes across UEP groups, all other populations are both significantly different from one another and from the Ashkenazic Israelites, Sephardic Israelites, and Yemeni ($P < .01$ in all cases).

Genetic Identity

Nei's genetic identity I, calculated on the basis of distributions of haplotypes at the UEP level, shows the Lemba to be closer to the four Semitic populations than to the Bantu.

Haplotypes

Table 2 lists the Y-chromosome haplotypes of all samples, classified by UEP group and microsatellites, identifying those Y chromosome haplotypes that are observed in more than one population. Only 2 (.014) of the 142 microsatellite haplotypes are duplicated across UEP groups, which provides some confidence that the six microsatellites used in this study are sufficient to prevent high levels of homoplasy, at least at fairly deep genealogical levels. Evolutionary convergence of the microsatellite haplotypes within UEP haplogroups will, of course, not be detected.

To evaluate sharing of Y chromosomes between populations, we considered both the total number of Y chromosomes observed (table 3, *top*) and a data set in which each Y-chromosome haplotype was counted only once (table 3, *bottom*). The degree of sharing of Y chromosomes between Ashkenazic and Sephardic Israelites has been reported elsewhere (Bradman et al., in press). Haplotype 34 (CMH) is present in varying frequencies in Ashkenazic Israelites, Sephardic Israelites, Yemeni, and Lemba but is absent from both the Bantu population and the Sena population. The CMH represents .088 of all Lemba Y chromosomes and .135 of Lemba UEP group 1 Y chromosomes; equivalent population (UEP group 1) proportions for Ashkenazic Israelites, Sephardic Israelites, and Yemeni are, respectively, .150 (.231), .100 (.161), and .020 (.027). The CMH is an important component in the sharing of Ashkenazic- and Sephardic-Israelite Y chromosomes. It is also the major component of Ashkenazic- and Sephardic-Israelite Y-chromosome sharing with the Lemba. The Yemeni-Hadramaut samples only have one such chromosome among them. Two striking features are the large proportion of haplotypes found in the Bantu that are also present in the Lemba (.448; table 3, *bottom*) and the even higher proportion of Y chromosomes (.740; table 3, *top*).

Gene Diversity

Table 4 gives gene diversity (Nei 1987) for all six populations, at the UEP-group level and at the 12-marker-haplotype level; the gene diversity is calculated separately for each UEP group, for all populations having ≥ 10 Y chromosomes within that group. Geographic isolation may be the reason why the Sena population is the most homogeneous at the all-haplotypes level.

Among the other five populations, the homogeneity of the Bantu is striking, particularly at the UEP-group level. Haplotype 117 and its one-step microsatellite neighbors comprise almost half the Bantu Y chromosomes and could be a candidate signature haplotype of the eastern Bantu expansion. It will be interesting to see whether the haplotype cluster is represented in other eastern Bantu, as well as in western Bantu and Cameroon Bantu populations. The two Wolof individuals shared the same haplotype (haplotype 128, a one-step neighbor of the Bantu modal haplotype) with each other and with two Pedi individuals, one Swazi individual, and one Ndebele individual.

The Bantu modal haplotype was not significantly as-

Table 2

Distribution of Y-Chromosome Haplotypes, across Six Populations

UEP Group and Haplotype Index Number	No. of Microsatellite Repeats in[a]						No. of Y Chromosomes in Population(s)						
	DYS19	DYS388	DYS390	DYS391	DYS392	DYS393	All	AI	SI	Y	S	L	B
YAP⁻GACCT													
1	13	12	22	10	13	13	1	1					
2	13	12	23	10	13	13	3		3				
3	13	12	24	11	13	13	1	1					
4	13	15	23	10	11	12	1		1				
5	13	15	24	10	11	12	1		1				
6	13	15	25	10	11	12	2		2				
7	13	19	23	10	11	12	1		1				
8	14	12	21	10	11	14	1			1			
9	14	12	21	10	11	15	4			1	3		
10	14	12	21	11	11	15	1			1			
11	14	12	22	10	14	11	1			1			
12	14	12	22	10	15	12	1					1	
13	14	12	23	10	13	13	4	3	1				
14	14	12	23	10	15	14	13					13	
15	14	12	23	11	15	14	5					5	
16	14	12	24	10	15	14	1					1	
17	14	12	24	11	14	13	1	1					
18	14	12	24	11	15	13	1					1	
19	14	12	24	11	15	14	2					2	
20	14	13	21	10	11	15	1			1			
21	14	13	23	10	11	12	2		1	1			
22	14	13	23	11	13	13	1		1				
23	14	13	25	10	11	12	1	1					
24	14	14	23	10	11	12	3				1	2	
25	14	14	23	10	11	13	1		1				
26	14	14	24	10	11	12	1		1				
27	14	15	22	10	11	12	2	1	1				
28	14	15	23	10	11	12	4		1	3			
29	14	15	23	10	12	12	1			1			
30	14	15	24	10	11	12	14			1		13	
31	14	15	24	11	11	12	1					1	
32	14	15	25	10	11	12	4	2	1			1	
33	14	16	23	9	11	12	2		1	1			
34	14	16	23	10	11	12	27	9	5	1		12	
35	14	16	23	10	11	13	3				3		
36	14	16	23	10	12	12	1		1				
37	14	16	23	11	11	12	4				4		
38	14	16	24	10	11	12	3	1		2			
39	14	16	24	10	13	12	16					16	
40	14	16	24	11	11	12	1			1			
41	14	16	25	10	11	12	5	1		4			
42	14	16	25	10	13	12	1	1					
43	14	16	26	10	13	12	1	1					
44	14	17	22	11	11	12	2			2			
45	14	17	23	10	11	12	6			5	1		
46	14	17	23	10	11	13	1			1			
47	14	17	23	11	11	12	12			1	11		
48	14	17	23	12	11	12	1				1		
49	14	17	24	10	11	12	2					2	
50	14	18	23	10	11	12	1			1			
51	15	10	21	10	11	13	1						1
52	15	10	24	10	11	13	6					1	5
53	15	10	24	11	11	13	2						2
54	15	11	20	10	10	13	1						1
55	15	12	21	10	11	15	1				1		
56	15	12	22	9	11	14	1		1				

(*continued*)

Table 2 Continued

UEP Group and Haplotype Index Number	No. of Microsatellite Repeats in[a]						No. of Y Chromosomes in Population(s)						
	DYS19	DYS388	DYS390	DYS391	DYS392	DYS393	All	AI	SI	Y	S	L	B
57	15	12	22	10	10	14	7					7	
58	15	12	22	10	11	12	3		1	1	1		
59	15	12	22	10	11	14	2		1	1			
60	15	12	23	10	11	13	5	5					
61	15	12	23	11	13	13	1	1					
62	15	12	24	10	11	13	1	1					
63	15	12	24	10	14	13	2		2				
64	15	14	23	11	12	13	1	1					
65	15	14	24	11	11	14	1			1			
66	15	15	23	10	11	12	1	1					
67	15	15	24	10	11	12	5				1	4	
68	15	15	24	11	11	12	5		1			4	
69	15	15	26	9	11	12	1			1			
70	15	16	22	10	11	13	1	1					
71	15	16	23	9	11	12	1	1					
72	15	16	24	9	11	12	2	1	1				
73	15	16	24	10	11	12	1			1			
74	15	16	24	11	11	12	1					1	
75	15	16	25	10	13	12	1	1					
76	15	17	23	10	12	12	1			1			
77	16	10	24	10	11	13	2						2
78	16	11	18	12	10	13	1						1
79	16	11	22	12	10	13	1						1
80	16	12	24	11	11	13	1	1					
81	16	13	22	10	10	12	1		1				
82	16	13	23	10	12	13	1	1					
83	16	14	24	10	11	12	1					1	
84	16	15	24	10	11	12	1	1					
85	17	10	24	10	11	13	1					1	
YAP$^-$GACTT:													
86	13	12	21	10	16	13	1		1				
87	13	12	22	10	15	13	3	3					
88	14	12	22	11	13	13	1	1					
89	14	12	23	10	10	14	5		5				
90	14	12	23	11	14	13	1			1			
91	14	12	24	10	13	12	2					2	
92	14	12	24	10	13	13	1		1				
93	14	12	24	10	14	12	5	2	2	1			
94	14	12	24	11	13	12	2	1	1				
95	14	12	24	11	13	13	2	2					
96	14	12	24	11	13	14	2		2				
97	14	12	24	11	14	12	2	1		1			
98	16	12	24	10	11	13	5	1		4			
99	16	12	26	11	11	13	1			1			
YAP$^+$AACCT:													
100	13	12	23	10	11	13	4		3	1			
101	13	12	23	10	12	13	1		1				
102	13	12	24	9	11	13	2	1		1			
103	13	12	24	10	11	13	7	2	1			4	
104	13	12	24	10	11	14	1	1					
105	13	12	24	11	11	14	2	1					1
106	13	12	25	9	11	13	2	2					
107	13	12	25	9	11	14	2	2					
108	13	13	24	9	11	13	1		1				
109	14	12	24	10	11	13	2	1	1				
110	14	12	24	10	11	14	1			1			
111	14	12	25	10	11	13	1						1

(*continued*)

Table 2 Continued

UEP Group and Haplotype Index Number	No. of Microsatellite Repeats in[a]						No. of Y Chromosomes in Population(s)						
	DYS19	DYS388	DYS390	DYS391	DYS392	DYS393	All	AI	SI	Y	S	L	B
YAP+AGCCT:													
112	14	12	21	10	11	14	1						1
113	15	12	20	11	12	13	3					3	
114	15	12	21	10	10	14	1					1	
115	15	12	21	10	10	15	6					6	
116	15	12	21	10	11	12	1					1	
117	15	12	21	10	11	13	22					6	16
118	15	12	21	10	11	14	9			1		1	7
119	15	12	21	10	11	15	6					4	2
120	15	12	21	10	12	15	1			1			
121	15	12	21	11	11	13	10					3	7
122	15	12	21	11	11	14	3					1	2
123	15	12	22	10	11	13	1						1
124	15	12	22	11	11	13	1						1
125	15	12	24	11	11	14	1					1	
126	16	12	21	10	9	14	1					1	
127	16	12	21	10	11	12	1					1	
128	16	12	21	10	11	13	8					2	6
129	16	12	21	10	11	14	6					2	4
130	16	12	21	10	11	15	4					1	3
131	16	12	21	10	12	14	1						1
132	16	12	21	10	12	15	2						2
133	16	12	21	11	11	13	2					1	1
134	16	12	22	10	11	14	1						1
135	16	12	22	10	11	15	1					1	
136	16	12	22	10	11	16	2					1	1
137	16	12	24	10	11	14	1					1	
138	17	12	21	10	11	13	2						2
139	17	12	21	10	11	14	3					1	2
140	17	12	21	10	11	15	2					1	1
141	17	12	24	10	12	13	1					1	
142	17	13	20	10	11	15	1						1
							399	60	50	49	27	136	77

Note.—Explanations of abbreviations are as in table 1.

[a] For nomenclature, see Kayser et al. (1997).

sociated ($P = .773$) with any single tribal group (Raymond and Rousset 1995). Furthermore, no particular association of tribe and haplotype was immediately apparent. Approximately 78% (60/77) of Y chromosomes were shared across tribes, whereas ~45% (13/29) of Y-chromosome haplotypes were similarly shared. The Lemba have no single dominant haplotype, but haplotypes 14, 30, 34, and 39 together represent more than one-third (~40% [$n = 54$]) of the Lemba total; furthermore, since they are all in UEP group 1, they represent >60% of that group.

Microsatellite Variability

As reported by Thomas et al. (1998), the distribution of alleles across UEP groups suggests that microsatellite DYS388 does not mutate in a consistent stepwise manner over its entire range of repeat counts. The distribution of DYS388 alleles in UEP group 1 is strikingly different from that in other UEP groups. UEP group 1 has a two-peak distribution in which most alleles are either ≤ 12 or ≥ 15, whereas in other UEP groups in this study its repeat length is never >13. It is possible that low- and high-repeat-number subgroups within UEP group 1 are genealogically distinct, but, with current data, it is not possible to exclude the possibility that frequent large deletions occur in high-repeat-number alleles, as has been observed in artificially constructed systems (Wierdl et al. 1997).

Kayser et al. (1997), analyzing European samples, reported a lack of variation at DYS388, with all samples typed having a low-repeat-number allele. All Bantu samples in the present study also had low-repeat-number DYS388 alleles, whereas Semitic populations have a

Table 3

Proportions of Y Chromosomes and Y-Chromosome Haplotypes Shared by Pairs of Populations

	AI	SI	Y	S	L	B
POPULATION	Y Chromosome					
AI	...	.367	.233	.000	.200	.017
SI	.280	...	.300	.020	.160	.000
Y	.286	.204	...	.163	.061	.020
S	.000	.037	.593	...	.074	.000
L	.125	.154	.191	.044	...	.184
B‘	.013	.000	.091	.000	.740	...
	Y-Chromosome Haplotype					
AI	...	.237	.184	.000	.079	.026
SI	.265	...	.235	.029	.118	.000
Y	.200	.229	...	.114	.088	.029
S	.000	.100	.400	...	.200	.000
L	.065	.087	.065	.043	...	.283
B	.034	.000	.034	.000	.448	...

NOTE.—Explanations of population abbreviations are as in table 1.

majority of Y chromosomes with high-repeat-number DYS388 alleles. The typing of additional populations may reveal whether high-repeat-number DYS388 alleles common in Near Eastern populations are found at high frequency elsewhere. For now, high-repeat-number DYS388 alleles appear to be diagnostic of the Near East. The distributions of alleles of the other five microsatellites are similar to the distributions reported by Kayser et al. (1997).

Frequently Encountered Haplotypes

Nine haplotypes are represented by Y chromosomes present at a frequency ⩾10% in one or more of the six populations. Figure 2 shows the populations in which they are found. The figure clearly illustrates the relationships uncovered in tables 1 and 3—namely, the apparent contributions of Bantu and Jews to the Lemba Y chromosomes, as well as the connection between the Yemeni and the Sena populations. It also highlights haplotype 39, which is private to the Lemba and constitutes >11% of the total Lemba Y chromosomes and 18% of the Lemba Y chromosomes in UEP group 1.

The Origins of Lemba Y Chromosomes

To explore with greater resolution the possible origins of Lemba Y chromosomes, genealogical trees were drawn that were based on microsatellite variation. These were used to assess, in the case of each Lemba haplotype, whether each has a close genealogical relationship with one or more haplotypes present in any of the other five populations.

Separate genealogical trees of the individual haplotypes were drawn for each UEP group by use of two distance measures: average squared distance (ASD) and proportion of shared alleles (Slatkin 1995; Goldstein and Pollock 1997). ASD trees for UEP group 1 haplotypes were drawn, with and without DYS388. As expected, given the high- versus low-repeat-number allele distribution, inclusion of DYS388 increased the internal structure of the UEP group 1 tree, with haplotypes containing high-repeat-number DYS388 alleles grouping within one part of the tree and with haplotypes with low-repeat-number DYS388 alleles grouping within the other part of the tree (fig. 3). The tree also shows the populations in which each haplotype is represented. As would be anticipated with a large sample of similar haplotypes, bootstrap values were usually low, with only 3/77 nodes recording values >60% after 500 resamplings.

Classification into UEP group and position in genealogical trees differentiated Semitic and Bantu haplotypes. The data in table 3 demonstrate that there is very little sharing of Bantu and Semitic Y chromosomes. The highest level, at 9%, is due to a single Bantu haplotype shared with the Yemeni population. Consistent with a deep genealogical division between the two groups, most Semitic and Bantu haplotypes could be clearly differentiated either by their membership in a UEP group associated with Semitic or Bantu haplotypes or by their positions in a microsatellite-based tree restricted to a single UEP group. In UEP group 1, for example, the Bantu Y chromosomes clustered within their own clade. Lemba Y chromosomes, on the other hand, were frequently closely associated with either Bantu or Semitic Y chromosomes but only very rarely with both of them simultaneously. Given that differences can be observed between Semitic and Bantu Y chromosomes, it is possible to suggest a Semitic or Bantu origin for Lemba Y chromosomes, on the basis of their genealogical proximity to Bantu and Semitic types.

In UEP group 1, the two Lemba Y chromosomes (haplotypes 52 and 85) that are members of an otherwise

Table 4

Gene Diversity at the UEP and 12-Marker Haplotype Levels

	GENE DIVERSITY IN					
	AI	SI	Y	S	L	B
UEP-group level	.516	.538	.428	.000	.480	.322
12-Marker-haplotype level:						
UEP group 1	.909	.937	.940	.779	.891	.781
UEP group 2	.826	.750	...	...	...	...
UEP group 3	.840	...	...	...	...	...
UEP group 4	...	...	...	...	.923	.884
Overall	.951	.956	.957	.779	.945	.919

NOTE.—Explanations of abbreviations are as in table 1. Gene diversity is not calculated for populations represented by <10 chromosomes.

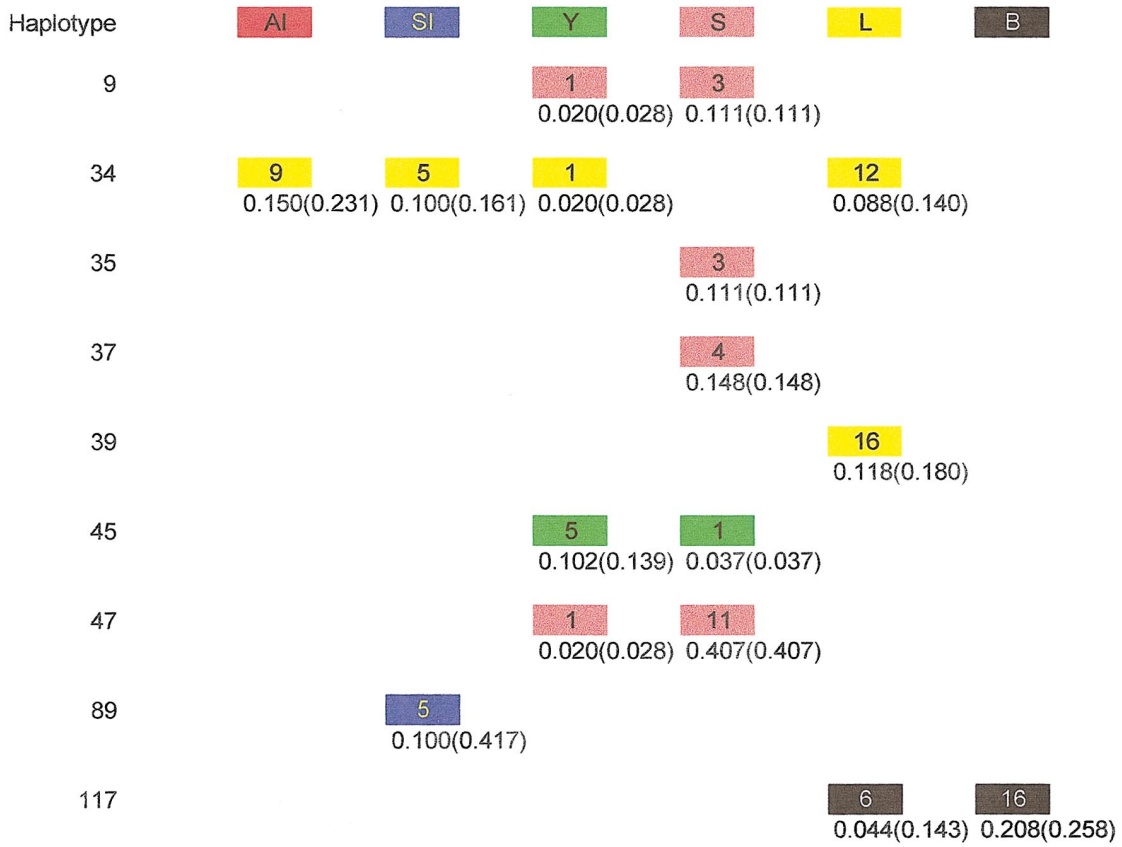

Figure 2 Frequently encountered haplotypes (frequency >10% in any population), in AI (*red*), SI (*blue*), Y (*green*), S (*pink*), L (*yellow*), and B (*black*); the colors indicate the population with the most chromosomes of the given haplotype. Numbers within the boxes are number of Y chromosomes in the population; numbers below the boxes are the frequency within the population (i.e., frequency within a UEP group). Explanations of abbreviations are as in table 1.

exclusively Bantu clade are classified as Bantu. Haplotype 57 Y chromosomes are classified as Semitic, since they are one-step neighbors of a Sephardic Israelite and a Yemeni, whereas the nearest member of the Bantu clade has an ASD distance six times greater. All other Lemba Y chromosomes in this haplogroup are classified as Semitic, since they are included within Semitic clades in the genealogical tree. Since all other Y chromosomes in UEP group 2 are from Semitic samples, the two Lemba Y chromosomes are classified as Semitic. The four Lemba Y chromosomes in UEP group 3 belong to a haplotype shared with three Israelites and are, therefore, classified as Semitic. Only 2/105 Y chromosomes in UEP group 4 are Semitic; both are Yemeni, and 1 of these 2 is shared with the Lemba and Bantu. Of the Bantu Y chromosomes not included within the Bantu clade in the UEP group 1 tree, 96.9% are included in UEP group 4. For these reasons, the Lemba Y chromosomes in UEP group 4 have been assigned a Bantu origin. The consequence of the above assignments is to designate 92 (67.6%) Lemba Y chromosomes as having a Semitic origin and to designate the other 44 (32.4%) of them as having a Bantu origin.

The Lemba Clans

For 108 of the 136 Lemba males sampled, clan affiliation was ascertained by self-identification. More than 10 samples per clan were collected from members of six clans (Buba, Hajji, Hamisi, Mhani, Sadiki, and Thouhakale). All six clans were represented in both the Northern Province and Sekhukuneland collections. Only the Mhani clan showed a significantly different distribution ($P < .02$) between the two locations, being more prevalent in the Northern Province collection. All six clans contained both UEP group 1 and UEP group 4 Y chromosomes, but none displayed UEP group 3 Y chromosomes. The Y-chromosome distribution among UEP groups was not significantly different ($P > .34$) between the two collection places.

Of particular interest is the Buba clan, since membership of this clan and possession of the CMH are

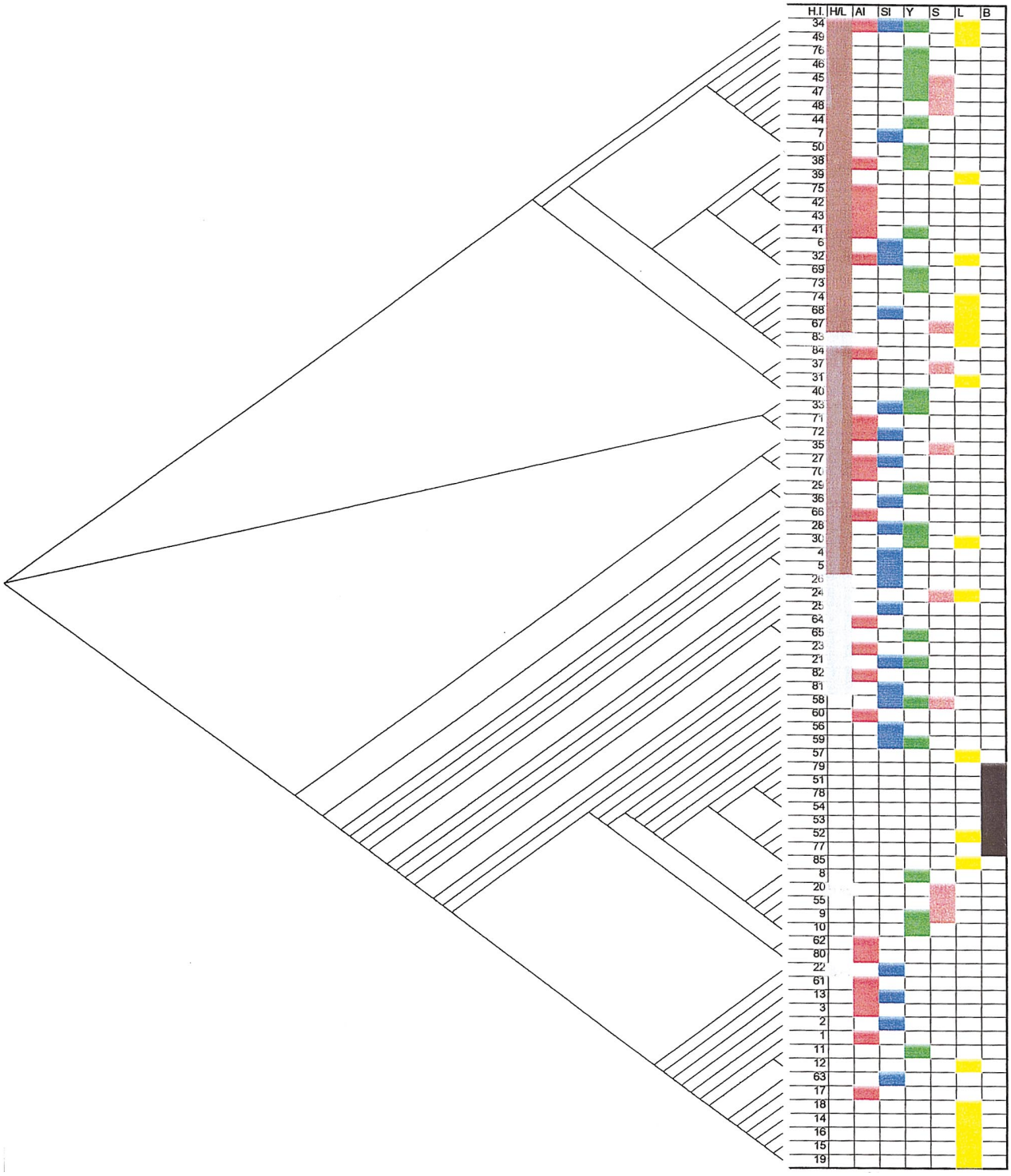

Figure 3 UEP group 1 genealogical tree of Y-chromosome haplotypes. The neighbor-joining tree was drawn by use of NEIGHBOR (part of the PHYLIP package), with the distance measure ASD calculated by use of the program MICROSAT. Colors next to the haplotype index numbers indicate the presence of Y chromosomes in the relevant population. H.I. = haplotype index number, H/L = high (≥15 [*brown*]), low (≤12 [*white*]), or intermediate (13 or 14 [*gray*]) number of repeats for microsatellite DYS388. Results are for AI (*red*), SI (*blue*), Y (*green*), S (*pink*), L (*yellow*), and B (*black*). Explanations of abbreviations are as in table 1.

significantly associated ($P < .0001$). Seven of the 11 clan-designated Lemba CMH Y chromosomes came from members of this clan, whereas 7 (Northern Province, 4/4; and Sekhukuneland, 3/9) of the 13 Buba have the CMH. F. C. Raulinga Hamisi, a Lemba elder, in a speech at the burial of Maanda William Mawela Ratshilingana Mhani, in July 1996 (before the current research was undertaken), said that "the Senas left Judea under the leadership of Buba and settled in Yemen where they built their city of Sena, hence Senas," reflecting the belief of at least one elder that Buba led the Lemba out of Judea. On the other hand, the *Encyclopedia Judaica* (1972) makes no mention of a Buba in Jewish history. In a book published privately in 1992, another Lemba elder wrote that "the Bhuba lineage came down from Judea as the leading lineage of the Basena when they left Judea in their early migration to the Yemen where they settled and built the city of Sena. They ruled over all the lineages in good manner" (Mathivha 1992, p. 23).

The Bantu Expansion

Bantu Y chromosomes are found in UEP group 1, UEP group 3, and, most frequently, in UEP group 4. The proportion of YAP^- Bantu Y chromosomes found in this study is consistent with the results of other investigators, who report a significant YAP^- element in the eastern Bantu (Hammer et al. 1998). The seven Bantu haplotypes, one of which is represented in the Lemba, and the Lemba haplotype 85 formed a distinct clade in trees based on use of both ASD and allele sharing as distance measures. Bootstrap support (45% in 500 resamplings of the ASD-based tree) is moderately good, given the large number of haplotypes in the tree and their generally close inferred genealogical relationships. It may be that the eight Y-chromosome haplotypes in this clade are representatives of a YAP^- genealogy present, in the eastern Bantu, alongside the dominant YAP^+ Y chromosomes. It will be interesting to see whether these or similar haplotypes are represented in other Bantu populations.

There is considerable archaeological and linguistic evidence to support an expansion of Bantu-speaking people throughout subequatorial Africa (Cavalli-Sforza et al. 1994). It is therefore interesting to see what dates are suggested by the level of microsatellite variance among the dominant, UEP group 4 Bantu Y chromosomes typed in this study. When there is good reason to believe that the correct ancestral Y chromosome can be identified—for example, by its presence at a high frequency (both alone and together with its one-step neighbors)—it is possible to use the method of Thomas et al. (1998), in which the time to coalescence to an ancestral Y chromosome is $ASD = \mu t$. A single-step mutation model is assumed, with μ being the average microsatellite mutation rate (per generation) and with t being the time (in generations). Haplotype 117, which is present across the Bantu tribes and in the Lemba, was selected as the ancestral Y chromosome. Alone it represents 25.8% of UEP group 4 Bantu-sample Y chromosomes, 58% together with its one-step neighbors. When the ASD for microsatellites DYS19, DYS390, DYS391, DYS392, and DYS393 is calculated and $\mu = .0021$ (Heyer et al. 1997) and 25 years/generation are used, the time to coalescence is 4,839 years. We also used the same approach as did Kittles et al. (Slatkin 1995; Kittles et al. 1998), to calculate an approximate time for the start of rapid population growth. In this method, $V = \mu t$, where V is the average variance of microsatellite-repeat counts. Using the same microsatellites, mutation rate, and generation time as used in the $ASD = \mu t$ calculation above produced a time of 3,310 years. Both times accord with estimates of the agricultural and Bantu expansions in subequatorial Africa (i.e., 3,000–5,000 years before the present [Cavalli-Sforza et al. 1994]). It should be appreciated that these dates carry very large confidence intervals, because of variability and uncertainty associated with (1) the evolutionary process and (2) the mutation rate and process. Since the estimates of confidence intervals depend on assumptions about demographic history (Goldstein et al. 1999), we have not attempted to estimate them here.

Discussion

Because of convergent evolution in microsatellites, care must be exercised in the interpretation of haplotype frequencies across populations. When two or more populations share haplotypes at high frequency, however, it is unlikely that convergence and matching drift is the correct explanation. The presence of Y-chromosome haplotypes in more than one population, and their absence from other populations, is suggestive of either common origin or male-mediated gene flow.

Genealogical trees of Y-chromosome microsatellite haplotypes included within the same UEP group classified by population can be very helpful in the identification of clades more associated with one population than with another—for example, the Bantu/Lemba clade in figure 3. When such population-associated clades are identified, it is possible that a search for UEP polymorphisms that distinguish between members of the identified clade and other members of the UEP group would yield biallelic markers useful in the study of relationships between particular population groups.

Sampling strategy in Y-chromosome population studies has not generally been given sufficient attention. Whether the collected samples truly reflect the structure of a population is difficult to assess. Since many populations may be both very heterogeneous and highly

structured geographically, researchers must be careful when extrapolating conclusions based on an analysis of samples from a restricted area. Care must also be exercised in the interpretation of data from samples that are of unknown provenance and for which only a broad description of origin is available.

The results reported above suggest a genetic history of the Lemba that is not incompatible with their oral tradition. Clearly, there has been a Semitic genetic contribution, including, quite probably, one from Arabs, given the Lembas' presence on the eastern coast of Africa, where Arabs have settled for centuries (Mathew 1963). Both Ashkenazic and Sephardic Israelites are geographically far removed from the Lemba, and, were it not for the Y-chromosome sharing between the Yemeni and Jewish populations, the occurrence of Jewish haplotypes in the Lemba population would be highly suggestive of gene flow between the two groups. However, given the extent of Y-chromosome sharing between the Yemeni and Jewish groups, the presence of such haplotypes because of gene flow from Arab sources cannot be discounted. Support for a Jewish contribution to the Lemba gene pool is, nevertheless, found in the presence, at high frequency in the Lemba, of the CMH (.088 of the entire population and .135 of UEP group 1); the CMH is also observed at moderate frequency in Ashkenazic Israelites (.150 and .231) and Sephardic Israelites (.100 and .161), but it was observed in only a single Yemeni (.020 and .028). Furthermore, in an unpublished study of Palestinian Arabs (A. Nebel, D. Filon, M. Faerman, A. Oppenheim, personal communication), the CMH was present at only very low frequency (<.025). The CMH has been suggested as a signature haplotype for the ancient Hebrew population, and it may be performing that function in this study (Thomas et al. 1998). Further support for Lemba oral history comes from the Buba/CMH association. However, it is possible that the Lemba CMH Y chromosomes are a consequence of a relatively recent event that, in Lemba oral tradition, has acquired a patina of antiquity.

The genetic evidence revealed in this study is consistent with both a Lemba history involving an origin in a Jewish population outside Africa and male-mediated gene flow from other Semitic immigrants (both of these populations could have formed founding groups for at least some of the Lemba clans) and with admixture with Bantu neighbors; all three groups are likely to have been contributors to the Lemba gene pool, and there is no need to present an Arab versus a Judaic contribution to that gene pool, since contributions from both are likely to have occurred. The CMH present in the Lemba could, however, have an exclusively Judaic origin.

The female contribution to the Lemba gene pool may be very different from the paternal, although still consistent with Lemba oral tradition. Soodyall (1993), analyzing mtDNA, found no evidence of Semitic admixture. Significantly, more than one-quarter of the Lemba sampled by Soodyall et al. (1996) had the African intergenic COII/$tRNA^{Lys}$ 9-bp deletion. Our study provides no evidence of a specific contribution from the ancestors of the present-day residents of Sena.

References

Bradman N, Thomas M, Goldstein DG. The genetic origins of Old Testament priests. In: Renfrew C (ed) Population specific polymorphisms. Cambridge University Press, Cambridge (in press)

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton, NJ

Encyclopedia Judaica (1972) Keter Publishing House, Jerusalem

Felsenstein J (1993) PHYLIP (phylogeny inference package) version 3.5c. Department of Genetics, University of Washington, Seattle

Goldstein DB, Pollock DD (1997) Launching microsatellites: a review of mutation processes and methods of phylogenetic inference. J Hered 88:335–342

Goldstein DB, Reich DE, Bradman N, Usher S, Seligsohn U, Peretz H (1999) Age estimates of two common mutations causing factor XI deficiency: recent genetic drift is not necessary for elevated disease incidence among Ashkenazi Jews. Am J Hum Genet 64:1071–1075

Hammer MF, Karafet T, Rasanayagam A, Wood ET, Altheide K, Jenkins T, Griffiths RC, et al (1998) Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. Mol Biol Evol 15:427–441

Hammer MF, Spurdle AB, Karafet T, Bonner MR, Wood ET, Novelletto A, Malaspina P (1997) The geographical distribution of human Y chromosome variation. Genetics 145: 787–805

Heyer E, Puymirat J, Dieltjes P, Bakker E, de Knijff P (1997) Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum Mol Genet 6:799–803

Kayser M, Caglia A, Corach D, Fretwell N, Gehrig C, Graziosi G, Heidorn F, et al (1997) Evaluation of Y-chromosomal STRs: a multicenter study. Int J Legal Med 110:125–133

Kittles RA, Perola M, Peltonen L, Bergen AW, Aragon RA, Virkkunen M, Linnoila M, et al (1998) Dual origins of Finns revealed by Y chromosome haplotype variation. Am J Hum Genet 62:1171–1179

Liesenbang G (1977) New light on Venda traditions: Mahumane's account of 1730. Hist Afr 4:162–181

Mandivenga EC (1983) Islam in Zimbabwe. Mambo Press, Gweru, Zimbabwe

Mathew G (1963) The east African coast until the coming of the Portuguese. In: Mathew RO, Mathew G (eds) The history of east Africa. Oxford University Press, Oxford, pp 94–127

Mathivha MER (1992) The Basena/Vamwenya/Balemba. Pub-

lished privately. Available from the author: P. O. B. 399, Shayandima, Thohoyandou, 0945, Republic of South Africa
Nei M (1987) Molecular evolutonary genetics. Columbia University Press, New York
Parfitt T (1997) Journey to the vanished city. Phoenix, London
Raymond M, Rousset F (1995) An exact test for population differentiation. Evolution 49:1280–1283
Ruwitah A (1997) Lost tribe, lost language? the invention of a false Remba identity. Zimbabwea 5:53–71
Skorecki K, Selig S, Blazer S, Bradman R, Bradman N, Warburton PJ, Ismajlowicz M, et al (1997) Y chromosomes of Jewish priests. Nature 385:32
Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457–462
Soodyall H (1993) Mitochondrial DNA polymorphisms in southern African populations. PhD thesis, University of the Witwatersrand, Johannesburg
Soodyall H, Vigilant L, Hill AV, Stoneking M, Jenkins T (1996) mtDNA control-region sequence variation suggests multiple independent origins of an "Asian-specific" 9-bp deletion in sub-Saharan Africans. Am J Hum Genet 58:595–608
Spurdle AB, Jenkins T (1996) The origins of the Lemba "black Jews" of southern Africa: evidence from p12F2 and other Y-chromosome markers. Am J Hum Genet 59: 1126–1133
Thomas MG, Skorecki K, Ben-Ami H, Parfitt T, Bradman N, Goldstein DB (1998) Origins of Old Testament priests. Nature 394:138–140
Wierdl M, Dominska M, Petes TD (1997) Microsatellite instability in yeast: dependence on the length of the microsatellite. Genetics 146:769–779