

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. Goldstein²

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

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

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 = cytosine, 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 $\geq .1$ 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-

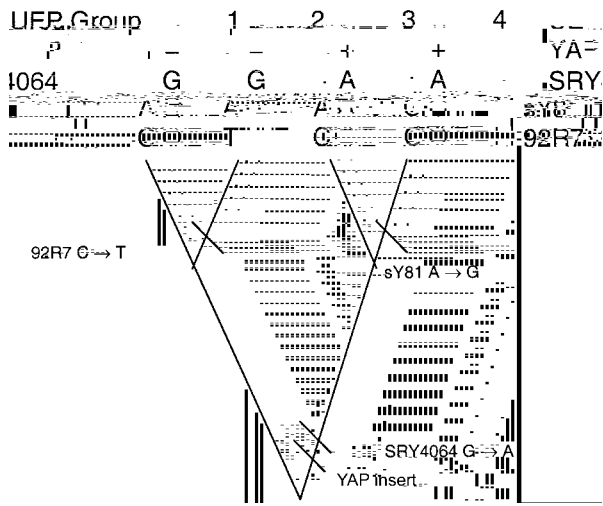


Figure 1

Table 2
Distribution of Y-Chromosome Haplotypes, across Six Populations

UEP GROUP AND HAPLO- TYPE 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 HAPLO- TYPE 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 3

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

	AI	SI	Y	S	L	B
POPULATION	Y Chromosome					
AI367	.233	.000	.200	.017
SI	.280300	.020	.160	.000
Y	.286	.204163	.061	.020
S	.000	.037	.593074	.000
L	.125	.154	.191	.044184
B'	.013	.000	.091	.000	.740	...
	Y-Chromosome Haplotype					
AI237	.184	.000	.079	.026
SI	.265235	.029	.118	.000
Y	.200	.229114	.088	.029
S	.000	.100	.400200	.000
L	.065	.087	.065	.043283
B	.034	.000	.034	.000	.448	...

NOTE

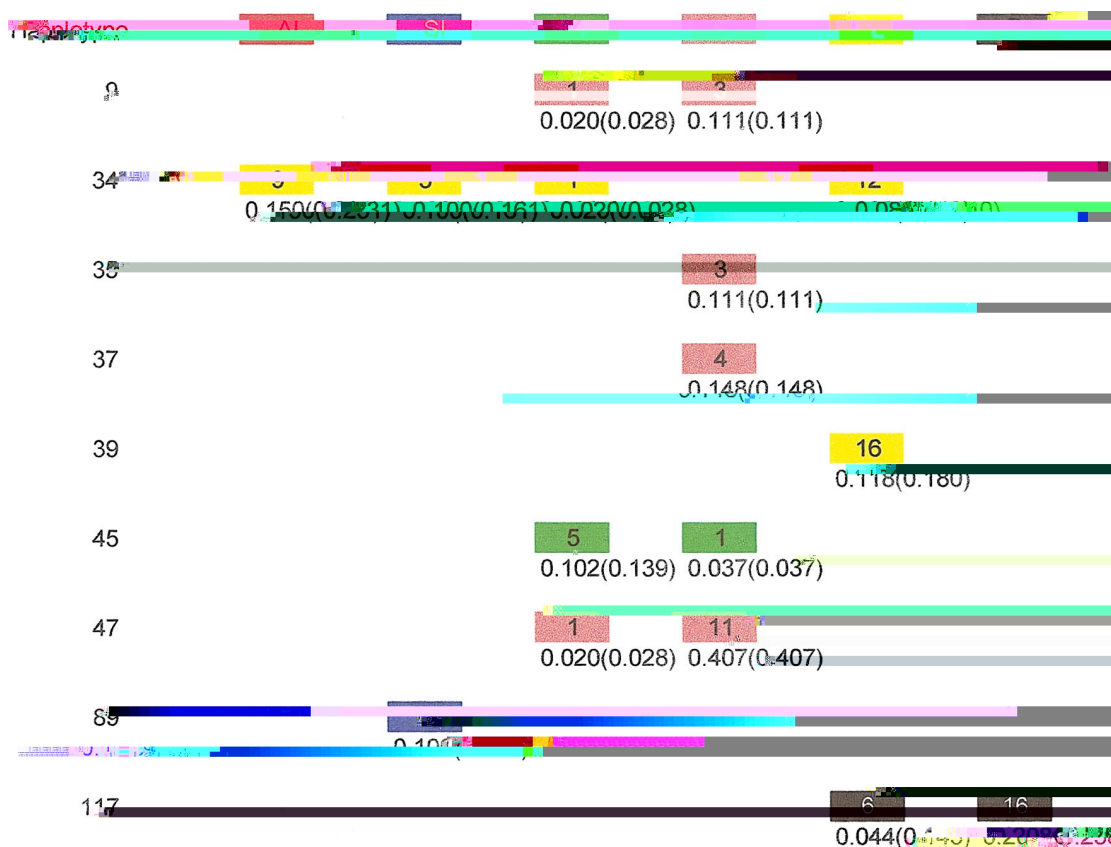


Figure 2 Frequently encountered haplotypes (frequency >10% in any population), in AI (red), SI (blue), Y (green), S (pink), L (yellow), and B (black); the colors indicate the population with the most chromosomes of the given haplotype. Numbers within the boxes are number of Y chromosomes in the population; 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 Bantuas8o27c; -395*[(grnon-(group)-43s312(wi5*[4(the)-456(most)4)-2abov5*[4(t-425(a0(atst)4)-

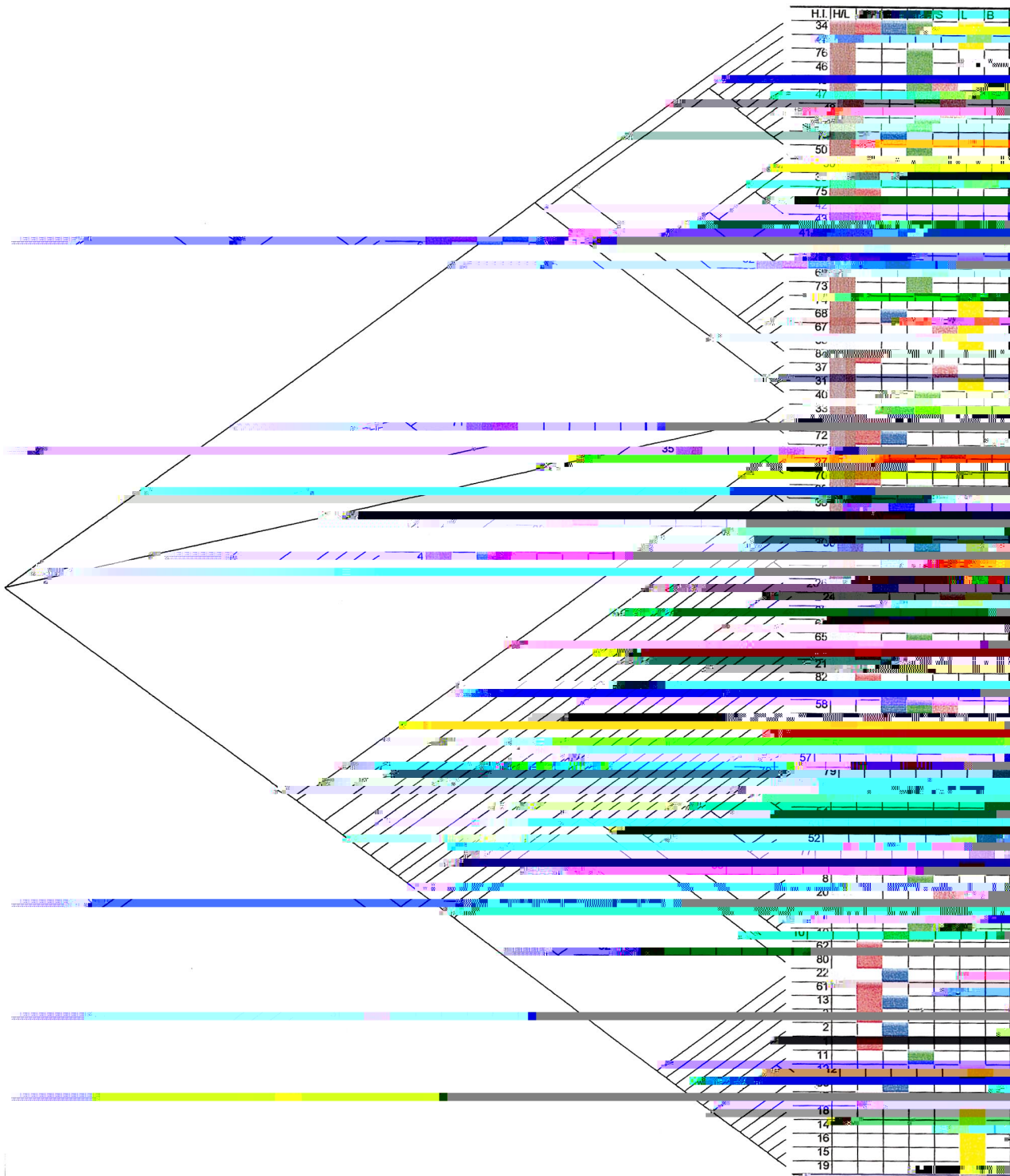


Figure 3 UEP group 1 genealogical tree of Y-chromosome haplotypes. The neighbor-joining tree was drawn by use of NEIGHBOR (part of the PHYLIP package), with the distance measure ASD calculated by use of the program MICROSAT. Colors next to the haplotype index numbers indicate the presence of Y chromosomes 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

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), an-

alyzing 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 HupslockB*(88:-33-342e)[TJT-1 -1.2222 TD[(Goldstein)-382(DB.)-3 Hc (1997)iesimsatsof mutatiors factfor genetic is for shkSenzzi J um, Gener , r,Rasanayagamf Wr, r., (1987) of Africg and aockrnays(is)-394(fe)-39((human)-39(Yn)-39(chfroom-)]TJT*[(omge)-344var , purdlge(r)118(.)--3(Bonneor)-233MRn, Wr, to, (1997)(the)--3((geograpicl.)-328adserim-)]TJT*[(bution)4222(of)-617(h (.)-350Dieljtts,de, (1997) omgespecific microsatellitl mutation um, Gene-6:7997261803m , ofo(1997) study39(.r)-344Iner(J)-350Legcl. ,,, ragon, , (1987) origins f reesaledomg(J)2367(um,)]TJT*-0.010 Tc[4(Gener)-28(62:1171-1179e)]TJ- umae's fe annivngaP (1893)Islamf if(Press,)]TJ1 -1.2222 TD[GwJer(.)-344Zimbabwoc ltewr (1693) the Africne fe ltewrlltewreds3) the of Afric,frsity-127e

- lished privately. Available from the author: P. O. B. 399, Shayandima, Thohoyandou, 0945, Republic of South Africa
- Nei M (1987) *Molecular evolutionary 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. *Zimbabwe* 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