European Journal of Human Genetics (2005), 1–10

www.nature.com/ejhg

ARTICLE

Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes

Elizabeth T Wood[1,2], Daryn A Stover[1], Christopher Ehret[3], Giovanni Destro-Bisol[4], Gabriella Spedini[4], Howard McLeod[5], Leslie Louie[6], Mike Bamshad[7], Beverly I Strassmann[8], Himla Soodyall[9] and Michael F Hammer*[,1,2]

[1]*Division of Biotechnology, University of Arizona, Tucson, AZ, USA;* [2]*Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA;* [3]*Department of History, University of Los Angeles, CA, USA;* [4]*Department of Animal and Human Biology, University La Sapienza, Rome, Italy;* [5]*Department of Medicine, Washington University, St Louis, MO, USA;* [6]*Children's Hospital of Oakland, Oakland, CA, USA;* [7]*University of Utah Health Sciences, Salt Lake City, UT, USA;* [8]*Department of Anthropology, University of Michigan, Ann Arbor, MI, USA;* [9]*Human Genomic Diversity and Disease Research Unit, University of Witwatersand, Johannesberg, South Africa*

To investigate associations between genetic, linguistic, and geographic variation in Africa, we type 50 Y chromosome SNPs in 1122 individuals from 40 populations representing African geographic and linguistic diversity. We compare these patterns of variation with those that emerge from a similar analysis of published mtDNA HVS1 sequences from 1918 individuals from 39 African populations. For the Y chromosome, Mantel tests reveal a strong partial correlation between genetic and linguistic distances ($r = 0.33$, $P = 0.001$) and no correlation between genetic and geographic distances ($r = -0.08$, $P > 0.10$). In contrast, mtDNA variation is weakly correlated with both language ($r = 0.16$, $P = 0.046$) and geography ($r = 0.17$, $P = 0.035$). AMOVA indicates that the amount of paternal among-group variation is much higher when populations are grouped by linguistics ($\Phi_{CT} = 0.21$) than by geography ($\Phi_{CT} = 0.06$). Levels of maternal genetic among-group variation are low for both linguistics and geography ($\Phi_{CT} = 0.03$ and 0.04, respectively). When Bantu speakers are removed from these analyses, the correlation with linguistic variation disappears for the Y chromosome and strengthens for mtDNA. These data suggest that patterns of differentiation and gene flow in Africa have differed for men and women in the recent evolutionary past. We infer that sex-biased rates of admixture and/or language borrowing between expanding Bantu farmers and local hunter-gatherers played an important role in influencing patterns of genetic variation during the spread of African agriculture in the last 4000 years.
European Journal of Human Genetics advance online publication, 27 April 2005; doi:10.1038/sj.ejhg.5201408

Keywords: mtDNA; Y chromosome; human; Africa; language; geography; correlation; evolution; Mantel

Introduction

An important question in population genetics is identifying the best predictors of genetic relationships among human populations. Several studies indicate strong correlations between genetic and linguistic relationships among globally distributed human populations.[1,2] At the subcon-

*Correspondence: Dr MF Hammer, Department of Ecology and Evolutionary Biology, Biosciences West, University of Arizona, Tucson, AZ 85721, USA. Tel: +1 520 621 9828; Fax: +1 520 621 9247;
E-mail: mfh@u.arizona.edu
Received 16 July 2004; revised 31 January 2005; accepted 15 February 2005

tinental scale, correlations between genetic variation and linguistic or geographic variation differ substantially. Y chromosome studies have shown that geographic distances correlate with genetic affinities among populations in Europe,[3] the Americas,[4] and Austronesia,[5] whereas language better explains Y chromosome relationships in Siberia.[6] Mitochondrial DNA (mtDNA) studies suggest that linguistic relationships are better correlated with genetic affinities among South American populations,[7] while both geography and language are correlated with maternal variation in Austronesia.[8] In Africa the question of gene–language relationships remains equivocal; some classical genetic and Y chromosome studies point to language,[9,10] while other Y chromosome and mtDNA studies identify geography[11,12] as a better predictor of genetic affinities.

The distribution of linguistic variation has been strongly influenced by the Neolithic Revolution, particularly in Africa. Linguistic, archeological, and ethnographic data suggest that all four African language families arose before agriculture in West Africa (Niger-Congo), Northeastern Africa (Afroasiatic), the middle Nile region (Nilo-Saharans), and East Africa (Khoisan).[13–17] Early dispersals of Niger-Congo, Afroasiatic, and possibly Nilo-Saharan languages are likely associated with migrating farmers.[14–17] Diamond and Bellwood[15] hypothesized that early farmers replaced the languages of hunter-gatherers living in their path of expansion and that this replacement would lead to strong correlations between linguistic and genetic variation. Their least equivocal example of an association of a language group with the spread of agriculture are the Bantu expansions. Beginning $\sim$4000 years ago, farmers speaking Niger-Congo Bantu languages expanded from a southern Cameroonian homeland over most of subequatorial Africa.[13,18,19] Evidence for the concordant spread of Bantu genes and languages comes from autosomal,[9,20] mtDNA,[12,21] and Y chromosomal[10,11,22–27] data.

The nonrecombining portion of the Y chromosome (NRY) and mtDNA are both haploid and uni-parentally inherited and, hence, are expected to have a four-fold reduction in effective population size (N_e) relative to the autosomes. In the absence of selection, the reduced N_e leads to an increased rate of genetic drift, which makes these haploid regions sensitive indicators of such demographic processes as bottlenecks, population subdivision, and population size and range expansions. The comparative study of patterns of variation at these loci allows the examination of the relative contribution of males and females in shaping African genetic diversity. In this study, we test for associations between genetic, linguistic and geographic differentiation to (1) identify correlates of genetic diversity in Africa, (2) examine the degree of concordance between the Y chromosome and mtDNA, and (3) assess the effects of sex-specific demographic processes shaping patterns of variation.

Subjects and methods

Population samples

Samples include representatives of the four major language families: Khoisan, Afroasiatic, Nilo-Saharan, and Niger-Congo (Table 1; Figure 1). Many of the 40 populations in Table 1 were analyzed in previously published studies;[22,23,28,29] however, several markers were typed in these samples for the first time in the current study. Differences in the number of samples in this and previous studies reflect differences in the availability of DNA, the inclusion of new samples, and/or the merging or splitting of populations according to language or ethnographic criteria. Sampling protocols were approved by the Human Subject Committee at the University of Arizona and those of collaborating institutions.

Y chromosome markers and terminology

Fifty biallelic Y-linked markers, SNPs and indels, were typed using a hierarchical protocol.[23,26,27] First, we typed mutations defining major haplogroups (eg, haplogroup A defined by M91) and then we typed all markers within a haplogroup until the most derived mutation in that haplogroup was determined (Figure 2). Thus, not every individual was typed for every marker. Markers were typed using allele-specific PCR, restriction enzyme digest, or direct sequencing. Protocols and primer sequences for these assays were previously published.[23,30] We follow the terminological conventions recommended by the Y Chromosome Consortium[30] for naming NRY lineages.

MtDNA

To compare maternally and paternally inherited patterns of variation, we re-examined 366 bp of mtDNA HVS1 sequence data compiled from a number of previous studies.[12,31–33] The data set includes 39 populations from the major language groups: Khoisan (!Kung1, !Kung2, Khwe, Hadza), Nilo-Saharan (Kanuri, Songhai, Turkana, Nubian, Sudanese, Mbuti, Datoga), Afroasiatic (Moroccan Berber, non-Berber Moroccan, Egyptian, Algerian Mozabite, Tuareg, Somalian, Amhara, Hausa, Podokwo, Mandara, Uldeme, Iraqw), Niger-Congo non-Bantu (Fulbe = Fulfulde, Yoruba, Serer, Wolof, Mandinka, Tupuri), and Niger-Congo Bantu (Bubi, Fang, Biaka, Kikuyu, Mozambique1, Mozambique2, Bakaka, Bassa, Mbenzele, Sukuma) (Figure 1). Some populations represented in the original data sets[12,31–33] were omitted because they are not found on the African mainland, are Cameroonian populations not represented in the Y chromosome data set,[33] or because linguistic designations could not be inferred.

Mantel tests

The correlation among genetic, linguistic, and geographic distances was assessed by the Mantel test[34] employing ARLEQUIN 2.000.[35] To test whether statistically significant associations between linguistic and genetic affiliations

Table 1 Sampled populations

Geographic region	*Ethnicity*	N	*Linguistic affiliation*		*Latitude/Longitude*
			Family	*Sublevel*	
West Africa					
Gambia/Senegal	Wolof	34	Niger-Congo	Atlantic	14N 15W
Gambia/Senegal	Mandinka	39	Niger-Congo	Atlantic	15:5N 15W
Mali	Dogon	55	Niger-Congo	Dogon	14N 3W
Ghana	Ewe	30	Niger-Congo	Kwa	6N 1E
Ghana	Ga	29	Niger-Congo	Kwa	5:5N 0E
Ghana	Fante	32	Niger-Congo	Kwa	6:5N 1:5E
Central Africa					
Cameroon, North	Podokwo	19	Afroasiatic	Chadic	12:5N 14:5E
Cameroon, North	Mandara	28	Afroasiatic	Chadic	12:5N 14:5E
Cameroon, North	Uldeme	13	Afroasiatic	Chadic	12:5N 14:5E
Cameroon, North	Tupuri	9	Niger-Congo	Gur	10N 13:5E
Cameroon, South	Bassa	11	Niger-Congo	Bantu	3N 10E
Cameroon, South	Bakaka	17	Niger-Congo	Bantu	3N 10:5E
Cameroon, South	Ngoumba	31	Niger-Congo	Bantu	3N 10E
Cameroon, South	Bakola Pygmies	33	Niger-Congo	Bantu	2:5N 10:5E
CAR	Biaka Pygmies	31	Niger-Congo	Bantu	3N 16E
CAR	Baka Pygmies	18	Niger-Congo	Bantu	3N 16E
East Africa					
DRC	Nande	18	Niger-Congo	Bantu	1:5N 30E
DRC	Hema	18	Niger-Congo	Bantu	1:5N 30E
DRC	Alur	9	Nilo-Saharan	Nilotic	1:5N 30E
DRC	Mbuti Pygmies	47	Nilo-Saharan	Sudanic	2N 29E
Tanzania	S. Cushitic	9	Afroasiatic	Cushitic	4S 35E
Kenya	Massai	26	Nilo-Saharan	Nilotic	2S 37E
Kenya	Luo	9	Nilo-Saharan	Nilotic	0:5S 34:5E
Kenya	Kikuyu & Kamba	42	Niger-Congo	Bantu	1:5S 38:5E
Uganda	Ganda	26	Niger-Congo	Bantu	1N 32E
Ethiopia	Amhara	18	Afroasiatic	Semitic	10N 39E
Ethiopia	Oromo	9	Afroasiatic	Cushitic	6N 39E
Ethiopia	South Semitic	20	Afroasiatic	Semitic	12N 38E
South Africa					
Namibia	!Kung/Sekele	32	Khoisan	Northern	19:07S 13:39E
Namibia	Tsumkwe San	29	Khoisan	Central	19:13S 17:42E
Namibia	Dama	18	Khoisan	Central	24:50S 17:00E
Namibia	Nama	11	Khoisan	Central	24:00S 17:50E
Namibia	Herero	24	Niger-Congo	Bantu	22:30S 18:58E
Namibia	Ambo	22	Niger-Congo	Bantu	22:30S 18:58E
South Africa	Sotho-Tswana	28	Niger-Congo	Bantu	27S 27E
South Africa	Zulu	29	Niger-Congo	Bantu	31:35S 28:47E
South Africa	Xhosa	80	Niger-Congo	Bantu	33:58S 25:36E
Zimbabwe	Shona	49	Niger-Congo	Bantu	17:50S 31:03E
North Africa					
Egypt	Egyptian	92	Afroasiatic	Erythraic	30N 30E
Tunisia	Tunisian	28	Afroasiatic	Semitic	35N 10E

reflect the same events in population history or parallel, but separate isolation by distance processes, we performed partial correlations holding geography (or language) constant.[36] Genetic distances were based on Slatkin's[37] linearized Φ_{ST} values (ie, incorporating molecular distances among haplogroups). Geographic distances between populations were calculated using approximate latitude and longitude data for the sample sites (Table 1). We used a novel approach for classifying linguistic relationships among populations. One of us (CE) constructed tree relationships among the languages spoken by the study populations using several sources of linguistic, archeological, and ethnographic data. Divergence times between related languages were estimated using archeological dates and glottochronological methods.[38] Linguistic relationships among populations in this study, as well as among the populations in the mtDNA data set, are available at http://www.u.arizona.edu/~ewood/data.html. We also

Figure 1 Map of Africa. The approximate location of 40 populations typed for Y chromosome markers in this study (●) and 39 populations surveyed for HVS1 sequence data[12,31,32,33] (○) are indicated. The distribution of the four African language families was constructed using Greenberg's[39] classifications and further refined with data from the ethnologue (http://www.ethnologue.com/). Three shades of gray on map refer to the distribution of language families: Khoisan (light gray, southwest), Afroasiatic (light gray, north), Niger-Congo (medium gray), and Nilo-Saharan (dark gray). The circled geographic regions include North, West, Central, East, and South Africa.

performed Mantel tests with matrices constructed using (1) the method described by Poloni *et al*[10] that uses the tree relationships of the languages defined by Greenberg,[39] (2) the tree relationships among languages reported in this study without making use of divergence times, (3) equal distances among populations of different language families, and/or (4) variable distances among populations of different language families. All matrices yielded very similar correlations (both *r* and *P* values) for the entire data set. Results differed slightly among matrices when we removed the Bantu speakers.

AMOVA

Analyses of molecular variance (AMOVA) were also performed using ARLEQUIN 2.000.[35] Both haplogroup frequencies and molecular differences among haplogroups were taken into account. We grouped populations by five geographic regions (West, Central, East, South, and North Africa) and by four linguistic groups (Afroasiatic, Nilo-Saharan, Khoisan, and Niger-Congo) (Figure 1). All samples used for the Mantel analysis were also used in the AMOVA: 1122 individuals from 40 populations for the Y chromosome and 1918 individuals from 39 populations for mtDNA.

Given that levels of population differentiation can be influenced by (1) sample composition and the Y chromosomal and mtDNA data sets presented here are sampled differently, (2) the differing rates and modes of evolution characterized by the Y chromosome and mtDNA systems, and (3) the type of polymorphisms examined (eg, preascertained Y chromosome SNPs *versus* mtDNA HVS1 sequence data), direct comparisons between these haploid

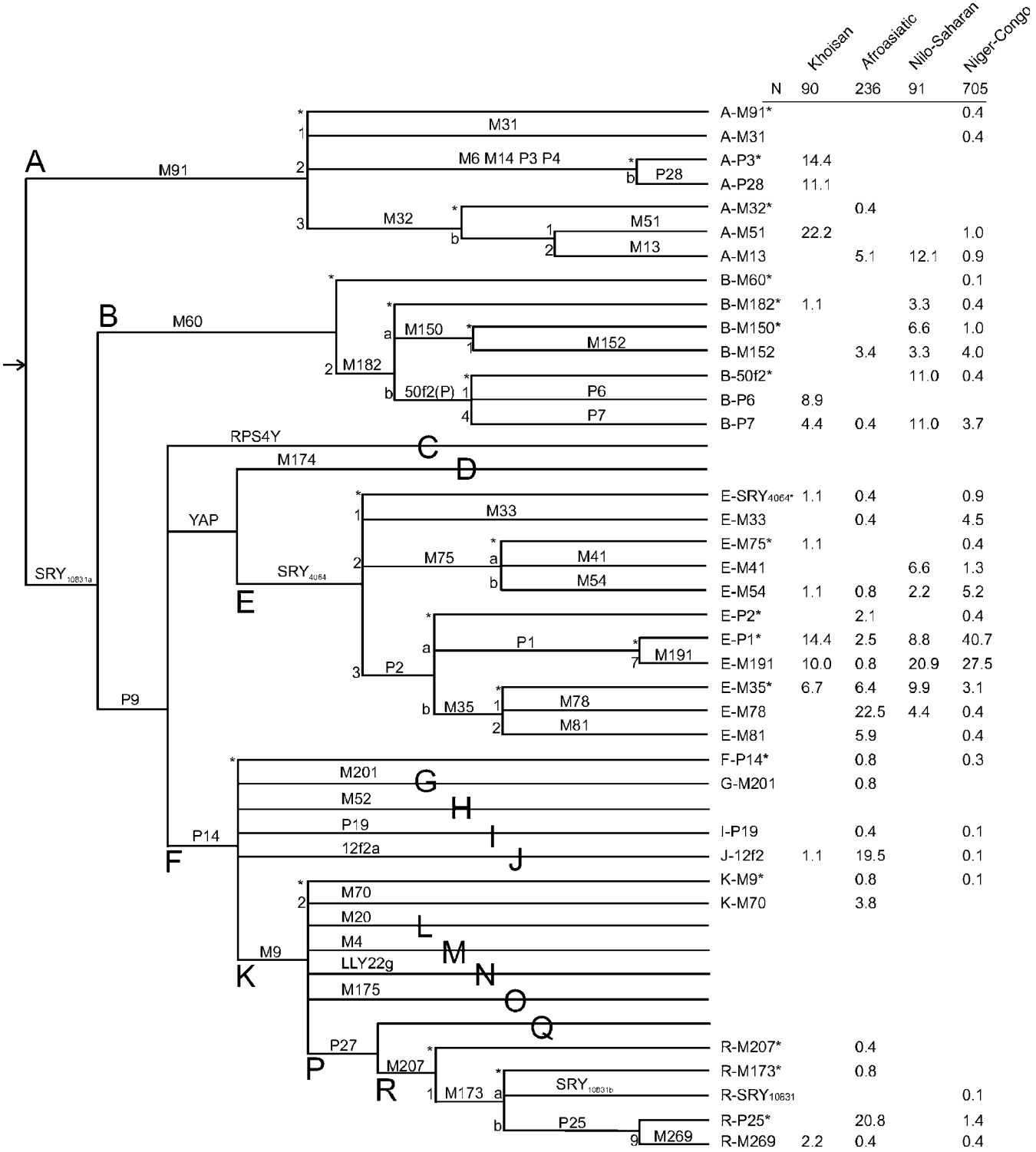

Figure 2 Maximum-parsimony tree of 50 Y chromosome biallelic markers typed in this survey. The root of the tree is denoted by an arrow. Major clades (ie, A–R) are labeled with large capital letters. Subclade labels (eg, A3b) are indicated to the left of the branches. Mutation names are given along the branches. The length of each branch is not proportional to the number of mutations or the age of the mutation. Only the names of the 36 haplogroups observed in the present study are shown to the right of the branches. Haplogroup frequencies are shown on the far right.

genetic systems should be considered with caution. Nevertheless, by comparing linguistic and geographic associations within a locus, we can ask whether mtDNA and the Y chromosome have been influenced by similar demographic processes.

Results

Geographic distribution of Y chromosome haplogroups in African populations

Phylogenetic analysis of the 50 Y biallelic markers used in this study yielded 36 haplogroups (Figure 2) (for appendix please refer to Supplementary Information). The vast majority of these lineages (98.1%) belong to five major haplogroups: A (7.1%), B (10.2%), E (70.2%), J (5.4%), and R (5.2%). Haplogroup A is closest to the root of the tree and is found most frequently in the Khoisan, particularly the A2 and A3b1 lineages (47.7%). Haplogroup B chromosomes are most frequently observed in Pygmies (48.9%), with B2a* and B2b* being nearly exclusive to this group. Haplogroup E is overwhelmingly the most common in this study. Over half of the individuals in our study (51%) are members of the subclade E3a, which is defined by the P1 mutation. Niger-Congo speakers have the highest frequency of E-P1* chromosomes (40.7%) and the largest proportion of E-M191 chromosomes (27.5%), particularly in Bantu speakers (31.5%). The E3b1 (E-M78) lineage is most frequent in Afroasiatics (22.5%). In this study, haplogroup J is concentrated in Afroasiatics (19.5%). While African haplogroup R chromosomes are generally quite rare, R-P25* chromosomes are found at remarkably high frequencies in northern Cameroon (60.7–94.7%). The remaining haplogroups (K, F*, I, and G) account for only 1.9% of the individuals in our data set.

Analysis of molecular variance (AMOVA)

The overall Y chromosome Φ_{ST} for the 40 populations is 0.32 (Table 2), a value that is similar to that found in a previous study of African Y-SNP diversity ($\Phi_{ST} = 0.34$).[28] This value is also similar to that obtained when our African sample is grouped into five geographic regions. When populations are grouped according to language family, the proportion of among-group variance ($\Phi_{CT} = 0.21$) is more than three times higher than when populations are grouped according to geographic location ($\Phi_{CT} = 0.06$) (Table 2). AMOVA results for the mtDNA data are also presented in Table 2. The continental mtDNA Φ_{ST} is 0.15. MtDNA Φ-statistics are very similar when populations are placed in either linguistic or geographic groups.

These results indicate that Y chromosome variation is significantly partitioned among both geographic and linguistic groups. Therefore, both language and, to a lesser extent, geography are probably important (albeit overlapping) predictors of African genetic structure.

Mantel tests

To test the underlying cause of association between genetic and linguistic *versus* geographic variation, we performed Mantel tests. These tests ask whether there is a correlation between geographic (or language) distance and genetic distance. Mantel tests reveal a statistically significant positive correlation between Y chromosome variation and linguistics ($r = 0.32$, $P = 0.001$) that explains 8.9% of the genetic variance. The correlation between genetic and linguistic variation remains strong when geography is held constant ($r = 0.33$, $P = 0.001$). In contrast, there is no correlation between paternal genetics and geographic distances ($r = 0.01$, $P > 0.10$) (Table 3). Mantel test results based on the mtDNA HVS1 data are also presented in Table 3. The correlation between maternal genetics and linguistics is significant ($r = 0.23$, $P = 0.016$), but weakens when geography is held constant ($r = 0.16$, $P = 0.046$). Similarly, a significant correlation between mtDNA and geography ($r = 0.23$, $P = 0.008$) weakens when linguistics is held constant ($r = 0.17$, $P = 0.035$). It is important to note that a failure to find correlations in Mantel tests does not mean that two variables are not related in some way. Rather, it means that processes that might cause a positive correlation (eg, isolation by distance or directional gene flow in the case of geography, or strict language–gene co-

Table 2 Analysis of molecular variance (AMOVA)

		Within populations		*Among populations within groups*		*Among groups*	
Group	*No. of groups*	*Variance (%)*	Φ_{ST}	*Variance (%)*	Φ_{SC}	*Variance (%)*	Φ_{CT}
Y chromosome	1	68.2	0.32				
Linguistic groups[a]	4	62.1	0.38	16.6	0.21	21.3	0.21
Geographic groups[b]	5	67.4	0.33	26.2	0.28	6.4	0.06
mtDNA	1	84.7	0.15				
Linguistic groups[a]	4	84.0	0.16	13.5	0.14	2.5	0.03
Geographic groups[b]	5	84.1	0.16	12.3	0.13	3.6	0.04

All Φ-statistics. *P*-values are less than 0.01.
[a]Afroasiatic, Khoisan, Niger-Congo, Nilo-Saharan (see Figure 1).
[b]West, Central, East, South, North Africa (see Figure 1).

Table 3 Correlation and partial correlation coefficients, *r* (*P*-value), between genetic, linguistic, and geographic distances

	Y	*mtDNA*
Genetics and Linguistics	0.32 (0.001)	0.23 (0.016)
Genetics and Linguistics, Geography held constant	0.33 (0.001)	0.16 (0.046)
Genetics and Geography	−0.01 (0.508)	0.23 (0.008)
Genetics and Geography, Linguistics held constant	−0.08 (0.859)	0.17 (0.035)

evolution) are unlikely to be the only processes operating (Tables 2 and 3).

Discussion

Mantel tests show a statistically significant positive correlation between Y chromosome and linguistic variation, while there is no correlation between Y chromosome and geographic variation. Furthermore, when populations are grouped according to language, the amount of among-group paternal differentiation (Φ_{CT}) is substantially higher than when grouped according to geographic location. Correlations with mtDNA show a different pattern. Maternal variation is weakly correlated with both language and geography and maternal among-group differentiation is nearly the same when populations are grouped according to linguistic affiliation or geographic location. These results suggest that patterns of differentiation and gene flow in Africa have been different for men and women in the recent evolutionary past.[10] In the following sections, we discuss (1) the relationships among genetic, linguistic, and geographic differentiation and the population history factors that may underlie these relationships, and (2) the effects of Bantu expansions on the distribution of Y chromosome and mtDNA variation in Africa.

Associations between genetic and linguistic variation

The association of genetic and linguistic variation has been observed at the global level,[1] as well as on the regional scale.[36,40,41] What are the underlying causes of these associations? Sokal[41] stated that a common language usually reflects a common origin for two populations, and a related language indicates a common origin farther back in time. This is generally thought to be an outcome of common historical processes leading to genetic and linguistic diversification – for example, a founding population may reproduce biologically and linguistically in a new location and replace the genes and languages of previous residents.[2,15,36] Discrepancies between genetic and linguistic differentiation could arise through a number of processes: genetic admixture can occur without language change, languages can be transmitted horizontally without significant genetic change, and/or genetic and linguistic evolution may proceed at heterogeneous rates.[2,15,42,43]

We found a statistically significant association between NRY variation and linguistic differentiation and a marginally significant association between mtDNA variation and linguisitic variation. However, when we performed Mantel tests controlling for geographic distance, the partial correlation between maternal genetic and linguistic variation weakens, while that between paternal genetic and linguistic variation remains statistically significant (Table 3). This suggests that the observed association between Y chromosome and language variation reflects the same co-evolutionary population history events.[38] These differing patterns for the Y chromosome and mtDNA could be the result of a greater degree of female than male admixture and/or the adoption of languages by females to a greater extent than males (see below). In either case, the implication is that African languages tend to be passed from father to children.[10]

Associations between genetic and geographic distances show the opposite trend than do the aforementioned associations between genetic and linguistic variation; there is no correlation between Y chromosome variation and geographic distance. In contrast, there is a stronger correlation between mtDNA variation and geographic distance, albeit only marginally significant when language is held constant (Table 3). Thus, the genetics–language correlation is stronger for the Y chromosome and the different pattern shown by mtDNA data suggests that men and women did not have identical demographic histories.

Effect of Bantu expansions on Y chromosome and mtDNA variation

Numerous studies suggest that the Bantu expansions have had a substantial impact on the distribution of genetic variation in Africa.[9,10,12,22,25,27] Is the strong Y chromosome–linguistics correlation we observe across the entire continent primarily the result of the massive migrations of Bantu farmers? We sought to clarify the effect of each language group on influencing the paternal genetics–linguistics association and the genetics–geography association by repeating the Mantel test after removing each language group in turn (ie, Afroasiatic, Khoisan, Nilo-Saharan, Niger-Congo non-Bantu, and Niger-Congo Bantu). If one language group were disproportionately contributing to the overall pattern, then the association is expected to weaken upon removal of this group. There is only a single language group that has this effect: the removal of Bantu–speakers causes the paternal genetics–linguistics correlation to drop from $r = 0.33$ to 0.08 (Figure 3).We note that the same trend is observed, albeit to a lesser extent, when other language matrices (see Subjects and Methods) are employed in the Mantel tests (data not shown). The lower correlation coefficient when Bantu populations (but not other linguistic groups) are

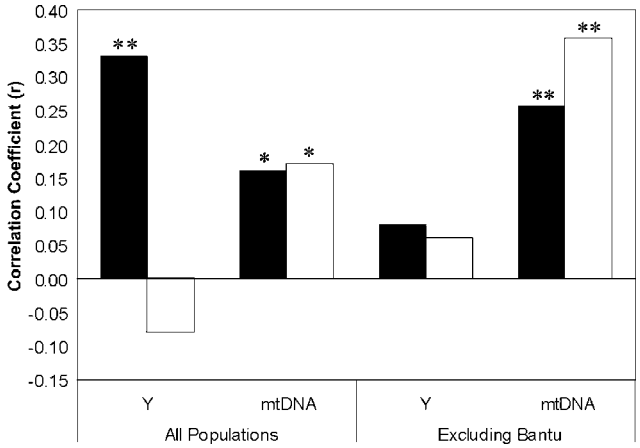

Figure 3 Partial correlation coefficient between genetics and linguistics holding geography constant (black bars) and genetics and geography holding linguistics constant (white bars) for the Y chromosome and mtDNA. $^{**}P<0.01$, $^{*}P<0.05$.

removed suggests that Bantu are contributing more to the language–Y chromosome relationship than any other language group.

The Y chromosome–geography correlation shows a different pattern. While the removal of the Bantu populations does not produce a correlation, the additional removal of four northern Cameroonian populations results in a statistically significant positive correlation ($r=0.27$, $P=0.008$). The strong effect of these northern Cameroonian populations on the Y chromosome results can be explained by the very high frequency of derived paternal (but not maternal) lineages that originated in non-African populations.[27,33] We note that this increased geographical correlation is not entirely attributable to the northern Cameroonian populations because when *only* these populations are removed, there is no Y chromosome–geography correlation ($r=-0.002$, $P>0.10$). Thus, in the absence of the unique populations from northern Cameroon, the removal of Bantu speakers leads to an association between Y chromosome and geographic differentiation, consistent with a recent dispersal of Bantu Y chromosomes.

On the other hand, the exclusion of Bantu speakers strengthens both the mtDNA–linguistics and mtDNA–geography correlations (Figure 3). Upon further exploration, we discovered that the increase in both of these maternal genetic correlations is due solely to Bantu-speaking Pygmy populations, specifically, the Biaka and Mbenzele (data not shown). The strong effect of these Pygmy populations on the mtDNA–linguistics correlation suggests that the horizontal transfer of languages from Bantu farmers to hunter-gatherer Pygmy females[19] occurred without significant genetic change.[19,31,36] The stronger mtDNA associations with both linguistic and geographic variation observed when the Biaka and Mbenzele populations are removed may reflect the fact that these two populations are maternal genetic outliers.[31]

To further investigate the effect of the Bantu expansions on patterns of geographic variation, we grouped populations by their geographic location and removed each language group in a series of four AMOVA runs (data not shown). Unlike the case for any other language family, the removal of Bantu populations results in a higher Y chromosome Φ_{CT} (0.28) than when they are included (0.06). This supports the hypothesis that Bantu Y chromosomes (eg E-P1*, E-M191) are acting to homogenize geographically differentiated populations. A similar analysis of mtDNA results in slightly higher Φ_{CT} value when the Bantu populations are excluded (0.07 *versus* 0.04).

If Bantu males and females dispersed equally from their West African homeland, replacing the genes of local hunter-gatherers in their path of expansion (equal sex ratio model), then we would expect similar patterns of association for paternally and maternally inherited loci. If, on the other hand, one sex dispersed more effectively (sex-biased model), we would expect to find differences in the degree of association between genetic and linguistic variation for the two haploid loci. Several explanations have been offered for observed differences in patterns of Y chromosome and mtDNA variation among populations.[31,44–46] Our results support the sex-biased model whereby the replacement of pre-existing languages by Bantu languages more closely parallels the turnover of Y chromosomes than mtDNA. How can this be explained? One possibility is that Bantu male farmers dispersed over longer distances or in greater numbers than Bantu females. Another possibility is that males and females dispersed equally, but there was a higher 'effective' migration rate for Bantu Y chromosomes than Bantu mtDNA. As Bantu farmers dispersed, they likely intermarried to some extent with the original inhabitants related to modern Pygmies and Khoisan.[15] In present-day African populations, the direction of intermarriage is usually between hunter-gatherer women and farmer (Bantu) men and the children of these marriages generally become farmers residing in their father's village[19,47] (ie, patrilocality). If this were typical of practices that existed throughout the Bantu expansions, we would expect Bantu mtDNA to be diluted (with hunter-gatherer mtDNA) to a greater extent than Y chromosomes. It is also possible that if the ancestral Bantu-speaking population were highly polygynous, then indigenous Y chromosomes would have been replaced by a more homogeneous pool of Bantu Y chromosomes, leading to a stronger correlation between linguistic and genetic variation. Indeed, polygyny is known to be substantially higher among food-producers than among hunter-gatherers.[31] In combination with the above processes, the adoption of Bantu languages by hunter-gatherer Pygmies may weaken maternal genetic–linguistic associations. Although our data cannot address the relative impact of

these sociocultural processes, it is likely that sex-biased migration/admixture, patrilocality, polygyny, and/or language borrowing contributed to the observed patterns of African variation.[31]

While earlier studies of Y chromosome variation have noted a correspondence between high-frequency haplogroups and the distribution of Bantu speakers (ie E-P1* and E-M191),[11,22–27] this is the first study to demonstrate a statistically significant correlation between Y chromosome SNP haplogroups[30] and linguistic differentiation in Africa. The data presented here are consistent with the hypothesis that prehistoric agriculture dispersed hand-in-hand with Bantu languages and Y chromosomes, with languages and Y chromosomes replacing those of hunter-gatherers in the paths of expansion.[15] Not all populations speaking Bantu languages in our study showed the effects of complete paternal genetic replacement (eg, the Bantu-speaking western Pygmies and northern Cameroonians). It is important to note that different mutation rates, as well as methods used to assay variation, on the NRY and mtDNA may contribute to some of the contrasting patterns observed here.[46] Future studies that examine Y chromosome and mitochondrial DNA sequence variation in the *same* samples representing African geographic and linguistic diversity will help to further elucidate the effects of Bantu expansions on the complex genetic landscape of Africa.

Acknowledgements

We would like to thank Alan Redd, Tanya Karafet, Leigh Hunnicutt, Roxane Bonner, and Jared Ragland for typing markers, and Matt Kaplan and the GATC for technical assistance. We also thank Brendan Hug, Phil Fischer, H Kimura, and AS Santachiara-Benerecetti for DNA samples, and Jason Wilder, Tanya Karafet, and Maya Metni Pilkington and three anonymous reviewers for helpful comments. Antonio Salas kindly provided a file with mtDNA data. GDB and GS were supported by MIUR (COFIN Grant No. 2003054059). This work was supported by grant GM53566 from the National Institute of General Medical Sciences to MH.

References

1 Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J: Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proc Natl Acad Sci USA* 1988; **85**: 6002–6006.
2 Chen JT, Sokal RR, Ruhlen M: Worldwide analysis of genetic and linguistic relationships of human-populations. *Hum Biol* 1995; **67**: 595–612.
3 Rosser ZH, Zerjal T, Hurles ME *et al*: Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am J Hum Genet* 2000; **67**: 1526–1543.
4 Zegura SL, Karafet TM, Zhivotovsky LA, Hammer MF: High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of Native American Y chromosomes into the Americas. *Mol Biol Evol* 2004; **21**: 164–175.
5 Hurles ME, Nicholson J, Bosch E, Renfrew C, Sykes BC, Jobling MA: Y chromosomal evidence for the origins of oceanic-speaking peoples. *Genetics* 2002; **160**: 289–303.
6 Karafet TM, Osipova LP, Gubina MA, Posukh OL, Zegura SL, Hammer MF: High levels of Y-chromosome differentiation among native Siberian populations and the genetic signature of a boreal hunter-gatherer way of life. *Hum Biol* 2002; **74**: 761–789.
7 Fagundes NJ, Bonatto SL, Callegari-Jacques SM, Salzano FM: Genetic, geographic, and linguistic variation among South American Indians: possible sex influence. *Am J Phys Anthropol* 2002; **117**: 68–78.
8 Lum JK, Cann RL, Martinson JJ, Jorde LB: Mitochondrial and nuclear genetic relationships among Pacific Island and Asian populations. *Am J Hum Genet* 1998; **63**: 613–624.
9 Excoffier L, Pellegrini B, Sanchez-Mazas A, Simon C, Langaney A: Genetics and history of sub-Saharan Africa. *Yrbk Phys Anthropol* 1987; **30**: 151–194.
10 Poloni ES, Semino O, Passarino G *et al*: Human genetic affinities for Y-chromosome P49a,f/TaqI haplotypes show strong correspondence with linguistics. *Am J Hum Genet* 1997; **61**: 1015–1035.
11 Scozzari R, Cruciani F, Santolamazza P *et al*: Combined use of biallelic and microsatellite Y-chromosome polymorphisms to infer affinities among African populations. *Am J Hum Genet* 1999; **65**: 829–846.
12 Salas A, Richards M, De la Fe T *et al*: The making of the African mtDNA landscape. *Am J Hum Genet* 2002; **71**: 1082–1111.
13 Greenberg J: Historical inferences from linguistic research in Sub-Saharan Africa; In: Butler J (ed): *Boston University Papers in African History I*. Boston, MA: Boston University Press, 1964, pp 1–15.
14 Ehret C: *The Civilizations of Africa: A History to 1800*. Virginia: The University Press of Virginia, 2002.
15 Diamond J, Bellwood P: Farmers and their languages: the first expansions. *Science* 2003; **300**: 597–603.
16 Ehret C: Historical/linguistic evidence for early African food production; In: Clark JD, Brandt S (eds): *From Hunters to Farmers*. Berkeley: University of California Press, 1984, pp 26–35.
17 Ehret C: Sudanic civilization; In: Adas M (ed): *Agricultural and Pastoral Societies in Ancient and Classical History*. Philadelphia: Temple University Press, 2001.
18 Ehret C: Linguistic inferences about early Bantu history; In: Posnansky CEaM (ed): *The Archaeological and Linguistic Reconstruction of African History*. Berkeley: University of California Press, 1982, pp 57–65.
19 Klieman K: *'The Pygmies Were Our Compass': Bantu and Batwa in the History of West Central Africa, Early Times to c. 1900 C.E.* Portsmouth, NH: Heinemann, 2003.
20 Cavalli-Sforza LL, Menozzi P, Piazza A: *The History and Geography of Human Genes*. Princeton, Princeton University Press, 1994.
21 Soodyall H, Vigilant L, Hill AV, Stoneking M, Jenkins T: mtDNA control-region sequence variation suggests multiple independent origins of an 'Asian-specific' 9-bp deletion in sub-Saharan Africans. *Am J Hum Genet* 1996; **58**: 595–608.
22 Hammer MF, Karafet T, Rasanayagam A *et al*: Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. *Mol Biol Evol* 1998; **15**: 427–441.
23 Hammer MF, Karafet TM, Redd AJ *et al*: Hierarchical patterns of global human Y-chromosome diversity. *Mol Biol Evol* 2001; **18**: 1189–1203.
24 Passarino G, Semino O, Quintana-Murci L, Excoffier L, Hammer M, Santachiara-Benerecetti AS: Different genetic components in the Ethiopian population, identified by mtDNA and Y-chromosome polymorphisms. *Am J Hum Genet* 1998; **62**: 420–434.
25 Thomas MG, Parfitt T, Weiss DA *et al*: Y chromosomes traveling south: the cohen modal haplotype and the origins of the Lemba – the 'Black Jews of Southern Africa'. *Am J Hum Genet* 2000; **66**: 674–686.
26 Underhill PA, Passarino G, Lin AA *et al*: The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann Hum Genet* 2001; **65**: 43–62.
27 Cruciani F, Santolamazza P, Shen PD *et al*: A back migration from Asia to sub-Saharan Africa is supported by high-resolution

analysis of human Y-chromosome haplotypes. *Am J Hum Genet* 2002; **70**: 1197–1214.
28 Hammer MF, Spurdle AB, Karafet T *et al*: The geographic distribution of human Y chromosome variation. *Genetics* 1997; **145**: 787–805.
29 Hammer MF, Redd AJ, Wood ET *et al*: Jewish and Middle Eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. *Proc Natl Acad Sci USA* 2000; **97**: 6769–6774.
30 YCC: A nomenclature system for the tree of Y chromosomal binary haplogroups. *Genome Res* 2002; **12**: 339–348.
31 Destro-Bisol G, Donati F, Coia V *et al*: Variation of female and male lineages in Sub-Saharan populations: the importance of sociocultural factors. *Mol Biol Evol* 2004; **21**: 1673–1682.
32 Knight A, Underhill PA, Mortensen HM *et al*: African Y chromosome and mtDNA divergence provides insight into the history of click languages (vol 13, pg 464, 2003). *Curr Biol* 2003; **13**: 464–473.
33 Coia V, Destro-Bisol G, Verginelli F *et al*: mtDNA variation in North Cameroon: lack of Asian lineages and implications for back migration from Asia to Sub-Saharan Africa. *Am J Phys Anthropol* 2005, in press.
34 Mantel N: The detection of disease clustering and a generalized regression approach. *Cancer Res* 1967; **27**: 209–220.
35 Schneider S, Roessli D, Excoffier L: *ARLEQUIN ver 2.000: A Software for Population Genetic Analysis*. Geneva, Switzerland: Genetics and Biometry Laboratory, University of Geneva, 2000.
36 Nettle D, Harriss L: Genetic and linguistic affinities between human populations in Eurasia and West Africa. *Hum Biol* 2003; **75**: 331–444.
37 Slatkin M: A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 1995; **139**: 457–462.
38 Embleton S: *Statistics in Historical Linguistics*. Bochum, Brockmeyer, 1986.
39 Greenberg JH: *The Languages of Africa*. The Hague: Mouton, 1963.
40 Barbujani G, Sokal RR: Zones of sharp genetic change in Europe are also linguistic boundaries. *Proc Natl Acad Sci USA* 1990; **87**: 1816–1819.
41 Sokal RR: Genetic, geographic, and linguistic distances in Europe. *Proc Natl Acad Sci USA* 1988; **85**: 1722–1726.
42 Cavalli-Sforza LL, Feldman MW: *Cultural Transmission and Evolution*. Princeton: Princeton University Press, 1981.
43 Barbujani G: DNA variation and language affinities. *Am J Hum Genet* 1997; **61**: 1011–1014.
44 Seielstad MT, Minch E, Cavalli-Sforza LL: Genetic evidence for a higher female migration rate in humans. *Nat Genet* 1998; **20**: 278–280.
45 Jorde LB, Watkins WS, Bamshad MJ *et al*: The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. *Am J Hum Genet* 2000; **66**: 979–988.
46 Wilder JA, Kingan SB, Mobasher Z, Pilkington MM, Hammer MF: Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females *versus* males. *Nat Genet* 2004; **36**: 1122–1125.
47 Jenkins T: *Human Evolution in Southern Africa: Human Genetics, Part A: The Unfolding Genome*. New York: Alan R. Liss, Inc., 1982, pp 227–253.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg).