Rare Deep-Rooting Y Chromosome Lineages in Humans: Lessons for Phylogeography

Michael E. Weale,* Tina Shah,*,1 Abigail L. Jones,* John Greenhalgh,* James F. Wilson,†
Pagbajabyn Nymadawa,‡ David Zeitlin,§ Bruce A. Connell,**
Neil Bradman* and Mark G. Thomas*,2

*The Centre for Genetic Anthropology, Department of Biology, University College, London WC1E 6BT, United Kingdom,
†Department of Biology, University College, London WC1E 6BT, United Kingdom, ‡Subassembly of Medical Sciences,
Mongolian Academy of Sciences, Ulaanbaatar-13, Mongolia, §Centre for Social Anthropology and Computing,
Department of Anthropology, University of Kent, Canterbury CT2 7NS, United Kingdom and
**Department of Languages, Literatures and Linguistics, York University,
Toronto, Ontario M3J 1P3, Canada

ABSTRACT

There has been considerable debate on the geographic origin of the human Y chromosome *Alu* polymorphism (YAP). Here we report a new, very rare deep-rooting haplogroup within the YAP clade, together with data on other deep-rooting YAP clades. The new haplogroup, found so far in only five Nigerians, is the least-derived YAP haplogroup according to currently known binary markers. However, because the interior branching order of the Y chromosome genealogical tree remains unknown, it is impossible to impute the origin of the YAP clade with certainty. We discuss the problems presented by rare deep-rooting lineages for Y chromosome phylogeography.

THE Y chromosome *Alu* polymorphism (YAP), first described by Hammer (1994), is a unique event polymorphism (UEP) marker defining a deep-rooting clade of the Y chromosome genealogical tree (Figure 1). Initial surveys showed that the YAP clade could be split into two main subgroups: one found only in East Asia and absent in Africa [here called group D following the nomenclature of the Y Chromosome Consortium (2002)] and one found mainly in Africa and absent in East Asia (here called group E; Spurdle *et al.* 1994; Hammer and Horai 1995; Hammer *et al.* 1997; Qian *et al.* 2000). Because group E represents the great majority of Y chromosomes found in sub-Saharan Africa [commonly found at local frequencies of between 65 and 100% (Cruciani *et al.* 2002)], there has been considerable interest and debate over the geographic origin of the YAP clade and the consequent implications for early human migration patterns. Hammer and colleagues used the position of group E within the (then) apparently paraphyletic group D to argue for an Asian origin of the YAP clade and a subsequent back-migration event that brought more derived YAP chromosomes to Africa from Asia (Altheide and Hammer 1997; Hammer *et al.* 1997, 1998, 2001). While this conclusion continues

to be cited (*e.g.*, Maca-Meyer *et al.* 2001; Templeton 2002), Underhill and colleagues have shown (through the discovery of the new marker M174: see Figure 1) that the Asian YAP subgroup is not paraphyletic and thus that the origin and direction of expansion of YAP chromosomes cannot be determined on these grounds (Underhill *et al.* 2000, 2001; Underhill and Roseman 2001).

Here we report a new very rare deep-rooting haplogroup within the YAP clade, together with data on other deep-rooting YAP clades (Figure 1). The new haplogroup, so far found only in five Nigerians, is the least derived of all YAP chromosomes according to currently known binary markers, such that application of the same phylogeographic inference method used by Hammer and colleagues (the nested cladistic method of Templeton *et al.* 1995) leads to the opposite conclusion—*i.e.*, significant evidence for range expansion from West Africa to Asia. However, we show that the apparently paraphyletic status of this haplogroup, and hence the conclusions of nested cladistic analysis, are also likely to be illusory. The interior branching order, and hence the origin, of YAP-derived haplogroups remains uncertain. We discuss the problems presented by rare deep-rooting lineages for Y chromosome phylogeography.

1*Present address:* Department of Medicine, University College, London WC1E 6JJ, United Kingdom.

2*Corresponding author:* The Centre for Genetic Anthropology, Department of Biology, University College, Gower St., London WC1E 6BT, United Kingdom. E-mail: m.thomas@ucl.ac.uk

NEW YAP LINEAGES

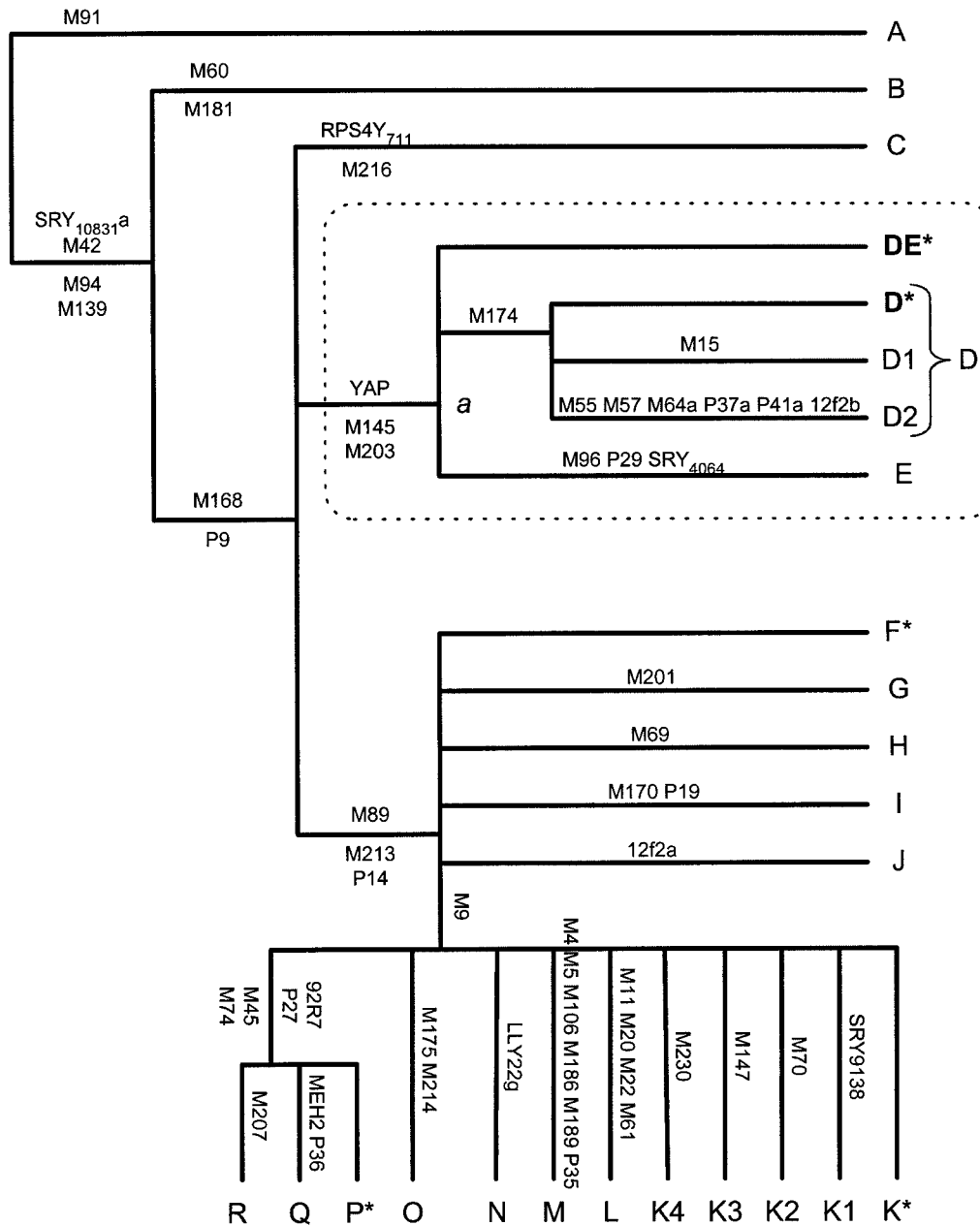Figure 1 depicts the YAP clade and deep-rooting subgroups within the Y chromosome genealogical tree as

FIGURE 1.—Y chromosome genealogical tree: major haplogroups as currently resolved by known UEP markers (indicated on respective branches). Dotted box indicates the YAP clade. Haplogroup D is shown split into three subgroups: D*, D1, and D2. Haplogroups DE* and D* are shown in boldface type. This is based on the published tree of the Y CHROMOSOME CONSORTIUM (2002) and updated with corrections and additions from KAYSER (2003) and KIVISILD (2003). Markers reflecting mutations that are known to have occurred more than once in the full Y chromosome genealogical tree (*e.g.*, 12f2) are shown with the suffix "a," "b," etc., to distinguish between the different mutation events.

described by currently known UEP markers. The new haplogroup, labeled DE* according to the nomenclature of the Y CHROMOSOME CONSORTIUM (2002), has been found in 5 Nigerians (from different villages, languages, ethnic backgrounds, and paternal birthplaces) from a data set of >8000 men worldwide, including 1247 Nigerians. The position of these 5 Nigerians on the Y chromosome tree has been confirmed by repeated typing for all the known UEP markers immediately above and below node *a* in Figure 1 (YAP, M145, M203, M174, M96, P29, and SRY$_{4064}$) as well as for five additional UEP markers (92R7, M9, M20, 12f2, and SRY$_{10831}$) as shown in Figure 1. The asterisk in DE* indicates that it is potentially, but not definitely, paraphyletic relative to one or both of groups D and E (Figure 2). The term

"paragroup" has been applied to such haplogroups (Y CHROMOSOME CONSORTIUM 2002). To help resolve the issue of paraphyletic status, we typed YAP-derived individuals in our data set for six microsatellites: DYS19, DYS388, DYS390, DYS391, DYS392, and DYS393. Of the five DE* individuals, three had a microsatellite haplotype consisting of repeat sizes 13-13-22-11-11-13 (loci arranged in same order as listed above) while the other two had a haplotype differing by one step at DYS391 only (13-13-22-10-11-13). This high level of similarity in such a rapidly evolving system strongly suggests that these five individuals share a private common ancestor (as in Figure 2, c, d, or e). We note that of the three possible branching patterns, two (Figure 2, c and d) would imply an African origin for YAP, while the third
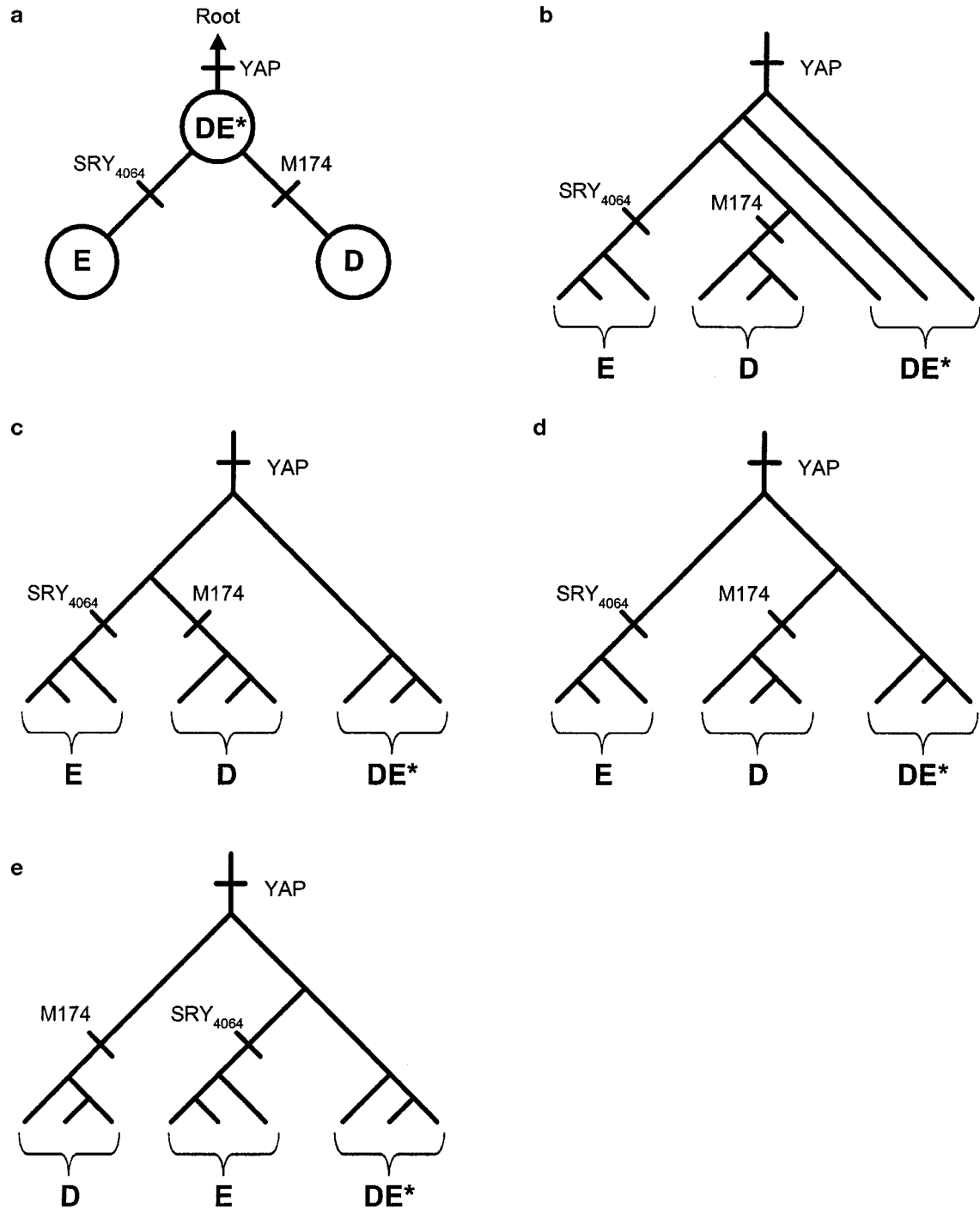
FIGURE 2.—Relationships of YAP+ haplogroups: (a) Cladogram linking D, E, and DE*. DE* is an "interior" haplogroup relative to the D and E "tip" haplogroups. (b–e) Different tree topologies that could give rise to the cladogram. (b) An example in which DE* is paraphyletic. Many other paraphyletic topologies are possible. (c) DE* is not paraphyletic but is still an outgroup to D and E. (d and e) DE* is neither paraphyletic nor an outgroup to D and E.

(Figure 2e) would leave the question of origins open. However, it is not easy to assess the relative probabilities of these three patterns.

In principle, the pattern of similarities of microsatellite haplotypes found in DE* and other YAP haplogroups could be used to deduce relative branching order. In practice, we found that no firm conclusions

could be made from a simple inspection of the microsatellite haplotype network, because haplotypes from different haplogroups were widely and evenly separated. A more formal BATWING analysis (WILSON *et al.* 2003; http://www.maths.abdn.ac.uk/ijw/downloads/download. htm, details available from M.E.W.) was also inconclusive. The D haplogroup was the most favored outgroup

(47% of posterior outcomes), but not to the exclusion of other possibilities (DE* outgroup in 24% of outcomes, E outgroup in 29% of outcomes). This uncertainty is increased even further by our uncertainty in the accuracy of the demographic and mutational model assumptions employed by BATWING. However, we note that, regardless of the relative branching order, the presence of the DE* haplogroup has the effect of forcing an earlier date for the most recent common ancestor of all African YAP chromosomes. This reduces the possible time window within which a back-migration to Africa could have occurred under the scenario of an Asian origin for YAP.

A new deep-rooting haplogroup (D* in Figure 1) has recently been reported by THANGARAJ *et al.* (2002) in a sample of 48 male Andaman and Nicobar Islanders taken from four different tribes. Remarkably, all 23 Onge and 4 Jarawa tribesmen were D*, while D* was completely absent in 10 Greater Andaman and 11 Nicobar tribesmen. Here we report the same haplogroup in two Mongolians (of Khalkh ethnic origin, the predominant ethnic group in Mongolia) from the same data set used to find the five Nigerian DE* chromosomes, which includes three population samples that contain group D individuals [422 Mongolians (of group D: 5 in D1 and 2 in D*), 77 Chinese (of group D: 17 in D1), and 38 Japanese (of group D: 19 in D2)]. These group D individuals have been typed for M174, M15, M55, and 12f2, as well as for the other markers typed in the DE* samples, but not for the four other mutations on the ancestral branch leading to group D2 (Figure 1). THANGARAJ *et al.* (2002) reported Y chromosome microsatellite haplotypes for 19 of the Onge D* individuals, including three loci (DYS19, DYS390, and DYS391) that overlap with ours. The two Mongolian individuals have very similar microsatellite haplotypes to each other (15-12-25-10-10-13 and 15-12-26-10-10-13, using the same ordering given for the DE* individuals). Likewise, the 19 Onge individuals share similar microsatellite haplotypes among themselves, but these differ markedly from the Mongolian individuals. This pattern strongly suggests a private common ancestor for the two Mongolian individuals and a separate private common ancestor for the Onge individuals. If this were the case, this would still leave the relative branching pattern of the four monophyletic groups within group D [D1, D2, D* (Mongolian), and D* (Onge)] unresolved [although BATWING analysis favored D2 as the outgroup among our D1, D2, and D* (Mongolian) samples in 59% of posterior outcomes]. Regardless of the branching order, however, the view that male Andaman Islanders descend from Asian colonizers is supported by these data.

IMPLICATIONS FOR PHYLOGEOGRAPHY

The exceptional haplotypic detail available on the Y chromosome, together with its high degree of geographic structure, has led to hopes that it represents the ideal tool for human phylogeographic analysis (*e.g.*, CRUCIANI *et al.* 2002). The new haplogroups reported here highlight some issues that stand in the way of this goal, however. The first of these is that it is easy to misinterpret apparently paraphyletic groups such as DE*. Inferences from nested cladistic analysis (TEMPLETON *et al.* 1995) depend crucially on the orientation of "interior" *vs.* "tip" haplogroups (haplogroups on a cladogram that appear, respectively, closer or farther from the assumed root). Figure 2 illustrates that the same cladogram can be created by many different underlying tree topologies. Conversely, if the "internal" haplogroup is not paraphyletic, then different cladograms, with different interior/tip orientations, can be created by the same tree by simple rearrangement of where the mutations are placed. This is because the only genealogically meaningful definition of the age of a clade is the time to its most recent common ancestor, but only if DE* is paraphyletic does it also become automatically older than D or E in this sense. Indeed, the very existence of observed interior haplogroups on a cladogram can be viewed as a missing-data problem: ideally, all the branches in the genealogical tree would be marked by different mutations at which point all observed haplogroups would appear as tips. It should also be noted that rejection of the null hypothesis (of no geographic structuring of clades) in nested cladistic analysis by no means guarantees the accuracy of the key used subsequently to infer the underlying process at work. When the predominant demographic forces at work are restricted gene flow or past geographic fragmentation, there are reasons to believe that signals inferred by nested cladistic analysis are more likely to be correct, because of an increased chance that an observed internal clade is also paraphyletic (see TEMPLETON 1998 and references therein). However, when past range expansions are the predominant factor, isolation can more easily result in cladograms where interior clades are monophyletic, as appears to be the case with DE*. TEMPLETON (1998) validated nested cladistic analysis empirically by applying the technique to 13 case studies (mitochondrial DNA data from several species) for which independent evidence existed for past range expansions, but the human Y chromosome may be a special case in this regard (see below).

Phylogeographic inferences based on parsimony reasoning are also open to misinterpretation. Prior to the discovery of M174 and of DE*, it was argued that an "African origin" hypothesis, involving one migration event (of ancestral YAP to Asia), one mutation event (SRY_{4064} to define group E), and one extinction event (ancestral YAP in Africa), should be considered as parsimonious as an "Asian origin" hypothesis, which also involves one mutation event (SRY_{4064}), one migration event (from Asia to Africa), and one extinction event (either group E in Asia or the pregroup E lineage in Africa, depending

on whether the SRY_{4064} mutation occurs before or after migration to Africa; ALTHEIDE and HAMMER 1997; BRAVI *et al.* 2000, 2001). However, from a genealogical point of view any two alternative migration events can be considered equally parsimonious only if they involve the migration of an equal number of ancestral lineages. In the above example, the number of ancestral lineages leading to group D was unknown at the time. As with nested cladistic analysis, the validity of inference depends on which underlying genealogy is correct.

Typing microsatellites on the Y chromosome is a useful way to try to resolve ambiguous cladograms into the underlying genealogical tree, but it may not always provide unequivocal results. In this case, we were able to infer that our DE* samples are likely to be monophyletic, but could not infer the relative branching order of D, E, and DE* with any certainty. It might be argued that because two of the three branching orders involve an African haplogroup as an outgroup (see Figure 2, c–e), there is a greater chance that this is the correct solution. Implicitly, this assumes that all branching orders are equally likely. However, since we already know that two of the haplogroups are African, it could be argued that this makes it more likely that the Asian haplogroup is the outgroup. In sum, the evidence weighs in favor of an African origin because two solutions (those with African outgroups) support it whereas the third (Asian outgroup) is only neutral. However, the strength of this evidence is unclear because the probabilities of the three solutions are unknown. Some attempt to infer these probabilities can be made using model-based analytical methods such as BATWING, but the success of such methods depends on the degree of temporal separation of the coalescence events of the deep-rooting branches leading to the different haplogroups and also depends on demographic and mutational assumptions that may be questionable.

The above arguments suggest that it is difficult, if not impossible, to make phylogeographic inference when the branching pattern is unresolved. Despite the large number of UEP markers known for the Y chromosome, with >200 currently described, many important deeply placed nodes are affected by this problem. The uncertainty in the branching orders of the 3 D, E, and DE* haplogroups or the 3 D1, D2, and D* haplogroups is minor compared to the uncertainty in the branching pattern of the 6 known M89-derived haplogroups or the 10 known M9-derived haplogroups (Figure 1), which together describe the majority of Y chromosomes found outside of Africa. Indeed, given the existence of potentially paraphyletic haplogroups at both these levels (F* and K*), the number of unresolved lineages may be even higher than this.

In the future, it is reasonable to expect that the uncertainties introduced by both unresolved paraphyletic status and unresolved branching order will be overcome by the discovery of new UEP markers from which the single, true, and unambiguous Y chromosome genealogy will emerge. Conservatively assuming a point mutation rate of 10^{-8}/nucleotide/generation on the 30-Mb euchromatic portion of the Y chromosome (BOHOSSIAN *et al.* 2000), one would expect a new mutation approximately once every three generations, so branching points separated by more than a few generations should be marked by a mutation somewhere. But even when the existing Y chromosome tree is completely resolved, the presence of rare deep-rooting lineages makes it reasonable to expect that more of such lineages are yet to be discovered and that existing ones may yet be found in other parts of the world. For example, in addition to the DE* and D* haplogroups reported here, WEALE *et al.* (2001) report a haplogroup (R1a1*) that is a paragroup to R1a1 (an important haplogroup in eastern Europe and in western, central, and southern Asia, defined by M17), which has been found only in 2 Armenian men from the same data set of >8000 men (including 734 Armenians) used here to find DE* and D*. In Figure 1, haplogroups such as K1 and K3 have also been identified only in very few individuals (KARAFET *et al.* 1999; UNDERHILL *et al.* 2000). Phylogeographic analyses can be highly influenced by the presence/absence status of different lineages in different geographical regions, and therefore our conclusions may change dramatically as new lineages are found.

Even if all extant deep-rooting lineages in the world are eventually found, and their branching order is characterized, the existence of rare deep-rooting haplogroups, together with the high geographic structuring of the Y chromosome, suggests a volatile birth-and-death process for Y chromosome lineages that also has implications for phylogeography. This may improve the usefulness of the Y chromosome in asking questions about recent or local human migration events, while at the same time introducing an unpredictable element into analyses of more ancient events. Some important deep-rooting lineages may die out completely, whereas others may expand and then contract to low frequency at some random location in its previous geographic distribution, confusing any inference on its point of origin. Uncertainty in the underlying demographic processes at work makes it difficult to compensate properly for this volatility by introducing population genetic modeling into phylogeographic analysis. For ancient phylogeography, the most promising way forward will be to compare results from several different unlinked loci, using the assumption that they act as independent replicates under the same demographic processes (TEMPLETON 2002). For inferences on male-mediated phylogeography alone, there will be limits to what the Y chromosome can tell us about ancient demographic events.

comments on an earlier draft of the manuscript, and all the donors who volunteered DNA for this study.

LITERATURE CITED

ALTHEIDE, T. K., and M. F. HAMMER, 1997   Evidence for a possible Asian origin of YAP(+) Y chromosomes. Am. J. Hum. Genet. **61:** 462–466.

BOHOSSIAN, H. B., H. SKALETSKY and D. C. PAGE, 2000   Unexpectedly similar rates of nucleotide substitution found in male and female hominids. Nature **406:** 622–625.

BRAVI, C. M., G. BAILLIET, V. L. MARTINEZ-MARIGNAC and N. O. BIANCHI, 2000   Origin of YAP+ lineages of the human Y-chromosome. Am. J. Phys. Anthropol. **112:** 149–158.

BRAVI, C. M., G. BAILLIET, V. L. MARTINEZ-MARIGNAC and N. O. BIANCHI, 2001   Tracing the origin and geographic distribution of an ancestral form of the modern human Y chromosome. Rev. Chil. Hist. Nat. **74:** 139–149.

CRUCIANI, F., P. SANTOLAMAZZA, P. D. SHEN, V. MACAULAY, P. MORAL *et al.*, 2002   A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. Am. J. Hum. Genet. **70:** 1197–1214.

HAMMER, M. F., 1994   A recent insertion of an Alu element on the Y-chromosome is a useful marker for human-population studies. Mol. Biol. Evol. **11:** 749–761.

HAMMER, M. F., and S. HORAI, 1995   Y-chromosomal DNA variation and the peopling of Japan. Am. J. Hum. Genet. **56:** 951–962.

HAMMER, M. F., A. B. SPURDLE, T. KARAFET, M. R. BONNER, E. T. WOOD *et al.*, 1997   The geographic distribution of human Y chromosome variation. Genetics **145:** 787–805.

HAMMER, M. F., T. KARAFET, A. RASANAYAGAM, E. T. WOOD, T. K. ALTHEIDE *et al.*, 1998   Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. Mol. Biol. Evol. **15:** 427–441.

HAMMER, M. F., T. KARAFET, A. J. REDD, H. JARJANAZI, S. SANTACHI-ARA-BENERECETTI *et al.*, 2001   Hierarchical patterns of global human Y-chromosome diversity. Mol. Biol. Evol. **18:** 1189–1203.

KARAFET, T. M., S. L. ZEGURA, O. POSUKH, L. OSIPOVA, A. BERGEN *et al.*, 1999   Ancestral Asian source(s) of New World Y-chromosome founder haplotypes. Am. J. Hum. Genet. **64:** 817–831.

KAYSER, M., S. BRAUER, G. WEISS, W. SCHIEFENHOVEL, P. UNDERHILL *et al.*, 2003   Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. Am. J. Hum. Genet. **72:** 281–302.

KIVISILD, T., S. ROOTSI, M. METSPALU, S. MASTANA, K. KALDMA *et al.*, 2003   The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. Am. J. Hum. Genet. **72:** 313–332.

MACA-MEYER, N., A. M. GONZÁLEZ, J. M. LARRUGE, C. FLORES and V. M. CABRERA, 2001   Major genomic mitochondrial lineages delineate early human expansions. BMC Genet. **2:** 13.

QIAN, Y. P., B. Z. QIAN, B. SU, J. K. YU, Y. H. KE *et al.*, 2000   Multiple origins of Tibetan Y chromosomes. Hum. Genet. **106:** 453–454.

SPURDLE, A. B., M. F. HAMMER and T. JENKINS, 1994   The Y-Alu polymorphism in southern African populations and its relationship to other Y-specific polymorphisms. Am. J. Hum. Genet. **54:** 319–330.

TEMPLETON, A. R., 1998   Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. Mol. Ecol. **7:** 381–397.

TEMPLETON, A. R., 2002   Out of Africa again and again. Nature **416:** 45–51.

TEMPLETON, A. R., E. ROUTMAN and C. A. PHILLIPS, 1995   Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum.* Genetics **140:** 767–782.

THANGARAJ, K., L. SINGH, A. G. REDDY, V. R. RAO, S. C. SEHGAL *et al.*, 2002   Genetic affinities of the Andaman Islanders, a vanishing population. Curr. Biol. **13:** 86–93.

UNDERHILL, P. A., and C. C. ROSEMAN, 2001   The case for an African rather than an Asian origin of the human Y-chromosome YAP insertion, pp. 43–56 in *Genetic, Linguistic and Archaeological Perspectives on Human Diversity in Southeast Asia: Recent Advances in Human Biology,* Vol. 8, edited by L. JIN, M. SEIELSTAD and C. XIAO. World Scientific Publishing, Singapore.

UNDERHILL, P. A., P. SHEN, A. A. LIN, L. JIN, G. PASSARINO *et al.*, 2000   Y chromosome sequence variation and the history of human populations. Nat. Genet. **26:** 358–361.

UNDERHILL, P. A., G. PASSARINO, A. A. LIN, P. SHEN, M. M. LAHR *et al.*, 2001   The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Ann. Hum. Genet. **65:** 43–62.

WEALE, M. E., L. YEPISKOPOSYAN, R. F. JAGER, N. HOVHANNISYAN, A. KHUDOYAN *et al.*, 2001   Armenian Y chromosome haplotypes reveal strong regional structure within a single ethno-national group. Hum. Genet. **109:** 659–674.

WILSON, I. J., M. E. WEALE and D. J. BALDING, 2003   Inferences from DNA data: population histories, evolutionary processes, and forensic match probabilities. J. R. Stat. Soc. A **166:** 155–201.

Y CHROMOSOME CONSORTIUM, 2002   A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res. **12:** 339–348.

Communicating editor: Z. YANG