Genetic diversity on the Comoros Islands shows early seafaring as a major determinant of human biocultural evolution in the Western Indian Ocean

Said Msaidie, Axel Ducourneau, Gilles Boëtsch, Guy Longepied, Kassim Papa, Claude Allibert, Ali Ahmed Yahaya, Jacques Chiaroni, Michael John Mitchell

Genetic diversity on the Comoros Islands shows early seafaring as major determinant of human biocultural evolution in the Western Indian Ocean.

Said Msaidie*[1], Axel Ducourneau*[1,2,5], Gilles Boetsch[1], Guy Longepied[2], Kassim Papa[1], Claude Allibert[4], Ali Ahmed Yahaya[3], Jacques Chiaroni[1] and Michael J. Mitchell[2]

[1] UMR 6578, Anthropologie Bio-culturelle, CNRS-EFS-Université de la Méditerranée, Faculté de médecine Hôpital Nord, Marseille, France.

[2] Inserm UMR S910, Génétique Médicale et Génomique Fonctionnelle, Inserm-Université de la Méditerranée, Marseille, France

[3] Laboratoire de Biologie de l'Hôpital El-Maarouf, Moroni, Comoros.

[4] Institut national des Langues et Civilisations orientales, Paris, France

[5] Leverhulme Centre for Human Evolutionary Studies, University of Cambridge, Cambridge, United Kingdom.

* SM and AD contributed equally to this work.

Running title: Genetic diversity in the Western Indian Ocean

Corresponding author :

Michael J Mitchell, Ph.D., Inserm UMR_S 910, Faculté de médecine, 27 bd Jean Moulin, 13385 Marseille cedex 05, France. Tel: +33 4 01257154; fax: +33 4 91804319.

e-mail: michael.mitchell@univmed.fr

Abstract

The Comoros Islands are situated off the coast of East Africa, at the northern entrance of the channel of Mozambique. Contemporary Comoros society displays linguistic, cultural and religious features that are indicators of interactions between African, Middle Eastern, and Southeast Asian populations. Influences came from the north, brought by the Arab and Persian traders whose maritime routes extended to Madagascar by 700-900 AD. Influences also came from the Far East, with the long-distance colonisation by Austronesian seafarers that reached Madagascar 1 500 years ago. Indeed, strong genetic evidence for a Southeast Asian, but not a Middle Eastern, contribution has been found on Madagascar, but no genetic trace of either migration has been shown to exist in mainland Africa. Studying genetic diversity on the Comoros Islands could therefore provide new insights into human movement in the Indian Ocean. Here, we describe Y chromosomal and mitochondrial genetic variation, in 577 Comorian islanders. We have defined 28 Y chromosomal, and 9 mitochondrial, lineages. We show the Comoros population to be a genetic mosaic, the result of tripartite gene flow from Africa, the Middle East and Southeast Asia. A distinctive profile of African haplogroups, shared with Madagascar, may be characteristic of coastal sub Saharan East Africa. Finally, the absence of any maternal contribution from Western Eurasia strongly implicates male-dominated trade and religion as the drivers of gene flow from the North. The Comoros provides a first view of the genetic makeup of coastal East Africa.

Keywords: Y chromosome, mitochondrion, Indian Ocean, East Africa, Comoros

Introduction

The Indian Ocean can be considered as a closed sea, an afro-Asiatic Mediterranean,[1,2] around which populations have migrated and mixed. In contrast to the Atlantic Ocean, which was a formidable natural barrier to East-West migration, the Indian Ocean with its seasonal monsoon winds favoured such exchanges, and most of the early trade routes were maritime. The Comoros archipelago is situated in the western Indian Ocean, midway between the island of Madagascar and the coast of East Africa at the northern end of the Mozambique Channel. The archipelago is composed of four main islands Grand Comore (*Ngazidja*), Anjouan (*Ndzuani*), Mohéli (*Mwali*) and Mayotte (*Maore*). The settlement of the four islands was an integral part of migration within the Indian Ocean, since they represent a potential maritime crossroads, and juncture, between Bantu African, Middle Eastern and Southeast Asian (SEA) spheres of influence. The modern Comorian population is the result of a long-term process of biocultural admixture, mainly related to ancient trade and colonisation in the Indian Ocean.

The Comoros and Madagascar share obvious signs of SEA influence including the cultivation of rice (phased out during 20th century), bananas and coconuts, and the use of outrigger canoes. Evidence from plant translocation suggests a migration from SEA 1 500 Years Before Present (YBP).[1,3,4] Clear genetic evidence for the SEA influence has been found on neighbouring Madagascar.[5-8] Based on Y chromosome and mitochondrial variation, ethnic groups with the strongest SEA biocultural features on Madagascar were estimated to have approximately 50% SEA ancestry.[5,8] In contrast to Madagascar where the language, Malagasy, is an Austronesian language with origins in SEA, the languages spoken on the Comoros are of Bantu origin. They are distinct from, but have close affinity to, Swahili, both branching from the precursor Sabaki language, 1 000 - 2 000 YBP.[9]

The cultural contributions of Middle Eastern civilisation are equally evident on the Islands. By 2 000 YBP, a thriving commercial maritime network already existed, extending from the Middle East to India, and as far South as Tanzania on the East African coast. The name "Comoros" is from the Arabic Kmr, meaning "light in the sky".[3] From 1 300 YBP, the Comoros archipelago served as a stepping-stone, for Middle Eastern traders operating along the East African coast, and for Southeast Asian traders travelling to Madagascar and the East African coast.[10,11] By 1 000 YBP, the Shirazi, traders with origins in the Persian city of Shiraz in present day Iran, had established themselves on the island of Kilwa. The Shirazi were responsible for the generalisation of Islam on the Swahili coast by 500 YBP. They had built mosques on Kilwa, Zanzibar and Anjouan by 800 YBP.[12] Islam remains the religion of the Islands today.

An unambiguous genetic signal from the Middle East has not, however, been detected in East Africa further south than Ethiopia,[13,14] or in the ethnic groups sampled on Madagascar.[5,8] The Lemba people of South Africa, carrying a putative semitic Y chromosome, currently provide the only evidence for gene flow from the Middle East into southern Africa.[15] The alleles of some autosomal genes found in the ex-patriot Comorian population living in Marseilles indicate a genetic contribution from Western Eurasia,[16-18] but the populations living on the Comoros have until now not been studied.

In this context, the peopling of the Comoros is evidently integral to the movements of men and women across the entire Indian Ocean. To gain insights into this process, we therefore determined the Y chromosomal and mitochondrial genetic variation on the three Bantu-speaking islands of the Comoros Republic.

Material and methods

Sample group

In February and March 2006, we obtained blood samples from 577 unrelated Comorian men (n=381) and women (n=196). We sampled the populations of three of the four islands of the Comoros archipelago (Grand Comore - 170 men, 67 women, Anjouan – 104 men, 69 women and Moheli – 107 men, 60 women). In 2006, this represented approximately 0.1% of the total Comoros population of 690 000 people. Blood was collected in EDTA vacutainer tubes and DNA extracted using the salting-out method.[19] Samples were collected from multiple towns and villages on each island (supplemental Figure 1). Recruitment was achieved through contacts established by medical personnel who originated from each community sampled. Each donor included in the study had four grandparents who were born on the same island and were native speakers of the island's language (Shingazidja, Shindzuani or Shimwali). Informed consent was obtained from all participants.

Y chromosome haplogroups and haplotypes

We typed 68 binary polymorphisms mainly by PCR-RFLP (Figure 1 and supplementary Table 1). For 293 Y chromosomes, alleles of 17 short tandem repeats (STR) polymorphisms on the Y chromosome were amplified with the AmpFlSTR Yfiler PCR Amplification Kit (Applied Biosystems). Y STR haplotypes were determined for 15 of 38 E-M2(xM191,U209) and 19 of 84 E-M191 chromosomes, and for all other chromosomes.

Mitochondrial haplogroups

We typed 31 coding region polymorphisms in the mitochondrial genome mainly by PCR RFLP (Figure 3, Supplementary Table 3). We also sequenced a 501 bp fragment including HVS-I from all M, N(xR) and R samples (Supplementary Table 4 and

GenBank: HM565257-HM565275). Choice of markers and branch designations was based on published data[20] and trees presented at http://www.ianlogan.co.uk.

Data analysis

The genetic structure (haplogroup number, haplogroup diversity, population differentiation, Fst and Rst) of the study population sample was analysed using the ARLEQUIN package v. 3.01,[21] and phylogenetic comparisons of Y STR haplotypes and mitochondrial SNP haplogroups were examined by Multidimensional Scaling (MDS) based, respectively, on Rst or Fst distances,[22,23] using SPSS 10.00 software. Admixture fractions were estimated using ADMIX 2.0.[24] Based on our haplogroup and MDS analyses, and historical and linguistic data, we chose Borneo, Iran and East Africa (Y: Kenya and Tanzania and mitochondrial: Mozambique) as the most likely parent populations from published data.[5,13,25-28]

Results

Y chromosome diversity on the Comoros

We analysed 381 Y chromosomes from the Comoros and identified 28 distinct haplogroups belonging to 11 of the 20 major clades of the Y chromosome tree as shown below (Figure 1).[29] These fall into four groups, based on the geographical distribution of haplogroups around the Indian Ocean: sub-Saharan African 59.6%; Western and Southern Asia 29.7%; Southeast Asia 6% and uncertain origin 4.7%. Four clades, E, J, O and R, have frequencies greater than 5% and represent 87.4% of the sample.

The paragroups C*(xC1-5), F*(xM282,M427), J* and K*(xLMNOPQRST), cannot be assigned an origin with certainty. Nevertheless the high frequencies of C*-M216 (Borneo – 2.5-25%) and K* (2-30%),[5,30] in SEA, make an SEA origin probable. J* has been found in Bali (1.5%) but also on the island of Soqotra (71%) situated in the

Gulf of Oman between Somalia and Yemen.[31,32] F*(xM282,M427) has been found mainly in the Indian subcontinent.[45] A West or Southwest Asian origin is therefore more likely for the F* and J* chromosomes.

Y STR analysis revealed a generally high variance (Table 1), which coupled with the large number of Y haplogroups, suggests that genetic drift has not drastically reduced genetic diversity on the Comoros Islands.

Sub-Saharan African Y chromosomes

The most common Comorian haplogroups, E1b1-M2 (41%) and E2-M90 (14%), are those that are frequent in sub-Saharan Africa.[13,33-36] They are present, respectively, at 56% and 6.4%, on Madagascar.[8] Two haplogroups were identified under E1b1-M2, derived for markers M191 (22%) and U209 (9%). The haplogroup E1b1a-M191 has been found in east and west sub-Saharan Africa, 19% in Tanzania and 57% in Benin.[13] The marker U209 was identified in Afro-Americans,[37] and has not, until now, been tested for in African populations.

The low incidence of E-M293 (0.8%) and A-M91 (0%) on the Comoros contrasts strongly with the frequency of these haplogroups in East African populations. E-M293 is found mainly in East Africa, Kenya and Tanzania (18%).[38] Furthermore, on the African mainland M293 chromosomes carry either 10, or 13 and more repeats at the DYS389I STR locus,[38] while, on the Comoros, they have 12 repeats. Haplogroup A has a frequency of 14% in Kenyan Bantu and 7% in Tanzania.[13]

Other haplogroups of likely sub-Saharan African origin on the Comoros are E-SRY_{4064}(xM2,M35,M75) (1.3%) and B2a (1.6%). B2a has a low frequency in southern Iran and Qatar,[27,39] but this is thought to be a consequence of the Arab slave trade. We therefore treat B2a as an African chromosome in this study.

Y chromosomes from around the Arabian Sea.

The northern Y chromosomes on the Comoros, E-V22, E-M123, F*(xF2, GHIJK), G2a, I, J1, J2, L1, Q1a3, R1*, R1a*, R1a1, and R2 (29.7%), make up a diverse group. G2a, J1 and J2 (16.5%) are thought to have originated in the Middle East.[14,40] J1-M267 has mainly spread south and west into the Arabic Peninsula, and into North and Northeast Africa, while J2-M172 lineages have expanded north into Europe and east into Asia.[13,14,39,41-43] The M78 subclade, E-V22, and E-M123 are believed to have originated in Northeast Africa, with E-V22 spreading to the west of North Africa and to the Arabic peninsula by the Levantine corridor (United Arab Emirates 6.7%),[39,44] while M123 spread mainly to the East (Yemen 8%; Oman 12%; Turkey 5.5%; Iran 1%).[13,27,39,40] In contrast the haplogroups L1, Q1a3, R1, R1a, R1a1 and R2 (10.5%), are thought to be of Central or Southern Asian origin and describe clines of decreasing frequency from India and Pakistan towards the Middle East.[45]

A comparison of the relative incidences of E-M78(V22), E-M123, G, J, L, Q and R on the Comoros with populations around the Arabian Sea shows greatest similarities with Southern Iran and, to a lesser extent, Turkey (Supplementary Figure 2).[27,40] The higher affinity to South Iran is also evident in the MDS analysis with the Comoros Y-STR data for the E-V22, E-M123, G, J, L Q and R haplogroups (Figure 2a). In the MDS, Comoros shows greatest affinity with UAE and South Iran. Southern Iran is the site of the first towns to develop in the Southern Middle East 2 000-3 000 years ago (Supplementary Figure 2).

A possible source of the Northern Y chromosomes is therefore the Shirazi traders from Southern Iran who established trading posts on the Comoros by 800 YBP.[12] It has previously been estimated that, at 9 Y-STR loci, 0-1 mutation will most

likely separate the descendants of a single Y chromosome haplotype after 40 generations (1 000 - 1 200 years).[46,47] Compatible with a Shirazi origin, we found that, at 9 Y-STR loci (DYS19, 389AB, 389CD, 390-393, 438 and 439), 42% of the Comoros Northern chromosomes differ by 0-1 mutation from chromosomes in Southern Iran.[48,49]

Southeast Asian Y chromosomes

We found the O1 lineage (6%) in the Comoros sample, providing genetic evidence for an SEA influence. Haplogroup O has been found at highest frequencies in East Asia and Island Southeast Asia.[50,51] All but one of the Comorian O1 chromosomes are O1a-M50 (5.8%). The O1a-M50 Y chromosome has its highest incidence in SEA: Borneo (10-20%), Sulawesi (4%), Taiwanese aborigines (0-59%, mean 14%) and the Philippines (3-12%).[5,52,53] It has not been detected in the Middle East or the Indian subcontinent.[5,27,39,45]

We performed an MDS with our STR data for the Y haplogroups O, C* and K* together with available STR data from candidate SEA populations (Figure 2b). The Comoros show a low affinity to the populations selected, even when C* and K* are not included (not shown), suggesting that these populations are not the source of SEA chromosomes on the Comoros.

Mitochondrial diversity on the Comoros

We have tested 577 Comorian samples for mitochondrial SNPs, and we define 9 distinct haplogroups (Figure 3). As for the Y chromosome, the majority of mitochondrial haplogroups on the Comoros, are of African origin. The haplogroups L0, L1, L2 and L3'4(xMN) compose 84.7% of the mitochondria in the Comoros sample, and their relative proportions, are most similar to profiles found in East and South East Africa.[20,54] The higher affinity with sub-Saharan East African populations is also

evident in the MDS analysis (Figure 4a and b).

The remaining 15.3% of the Comoros sample is composed almost exclusively of haplogroups that can either be unambiguously identified as SEA (B4a1a1-PM, F3b, and M7c1c - 10.6%),[25] or fall into the paragroup M(xD,E,M1,M2,M7) (4%) (Figure 3). The latter haplogroups are probably also originally from Southeast Asia, but of the 12 different M* HVS-I sequences on the Comoros, only two match published sequences: two M(xM7) mitochondria found on Madagascar.[8] We found no haplogroups that could be assigned to the Middle East.

Southeast Asian mitochondria

Of the SEA mitochondrial haplogroups present on the Comoros, F3b and M7c1c, like the Y-Hg O1-M50, each define an area of distribution that extends from Taiwan through the Philippines to Borneo[25] (Supplemental Figure 3). The MDS analysis with SEA populations shows greatest affinity to the Philippines and Borneo, although affinity is relatively weak (Figure 4c). Linguistic studies indicate Southeast Borneo to be the probable origin of the migration from SEA to Madagascar.[55] B4a1a1-PM (0.7%) is the major haplogroup throughout Polynesia (78%),[30] and on Madagascar (25%), but, within island SEA, it has not been found further West than South Borneo (1%).[5,7,8]

Male-biased gene flow from the Middle East

There are no mitochondrial lineages on the Comoros that are frequent in the Middle East (Figure 3). We have tested for, but did not find, the R haplogroups, H, J, T, U and V, or N(xR) that represent 80% of the mitochondria in Iran.[56] There is therefore striking evidence for male-biased gene flow from the Middle East to the Comoros, even if the unassigned mt-Hg M* and R* are designated as western Asian: 103/381 Y *vs*

27/577 mitochondria - Fisher's exact test, one sided, $p<10^{-22}$. This is entirely consistent with male-dominated trade and religious proselytisation being the forces that drove the Middle Eastern gene flow to the Comoros. For African and SEA contributions, if Y haplogroups C* and K* are counted as SEA, the under representation of the male lineages are similar (Y to mt ratio: Africa 0.69, SEA 0.66).

An opposite female gene flow from Africa to the Middle East, is clearly evident in Yemen (34% mt-Hg L; 4% Y-Hg E-M2), Iraq and the Levant.[57] However, no mt-Hg L has been found in Iran (n=712),[56] despite the presence of Y-Hg E-M2 (1.7%),[27] supporting the idea that the elevated mt-Hg L frequency in the western Middle East is not exclusively a consequence of the Arab slave trade, but also of geography.[58]

Discussion

We reveal the Comoros population to be a genetic mosaic, the result of tripartite gene flow from the North, the East and the West. Admixture analysis of the maternal and paternal contributions indicates the gene pool to be predominately African (72%), with significant contributions from Western Asia (17%) and Southeast Asia (11%). Our study therefore provides the first unequivocal evidence that the Middle Eastern trade routes that developed along the East African coast, during the last 2 000 years, have left a genetic trace. Male and female SEA gene flow has already been described on Madagascar, in populations that speak Austronesian languages,[5,8] but here we show that this extends beyond Madagascar, into African populations speaking languages from the Bantu family. This raises the question of whether the demic migration from SEA reached the East African mainland.

The frequencies of Y-Hg E-V22, E-M123, G2a, J1, J2, R1a1 and R2 in the Comorian sample are compatible with gene flow from Iran.[27,45] This concords with

historical data which attests to the presence of traders from Shiraz in Iran on the Comoros, and also the Comorian's own oral traditions which recount that Shirazi princes came in ships and established colonies on the islands. On the island of Anjouan the term "Shirazi" is used to refer to someone of Middle Eastern appearance. There is historical evidence that 1 000 YBP Persian traders played an important role in trade along the East coast, and we therefore predict that an Iranian genetic signal will be detected among Swahili speakers at former Middle Eastern trading centres on the sub-Saharan East coast, such as the islands of Zanzibar and Kilwa off the coast of Tanzania.

Interestingly, there are a number of similarities between the genetic profile of the Comoros islanders and the Lemba of South Africa, a Bantu speaking people whose Semitic origins are evident at both the cultural and genetic level.[15,59] The Lemba have high frequencies of the Middle Eastern Y chromosome HgJ-12f2a (25%), a potentially SEA Y, Hg-K(xPQR) (32%) and a Bantu Y, E-PN1 (30%) (similar to E-M2), raising the possibility that the Lemba and Comorian populations are consequences of similar demographic processes. The high-resolution genotyping of the Lemba Y chromosomes and mitochondria will elucidate this question.

The Comoros and Madagascar show similarities in the paternal and maternal contribution from SEA and Africa. The absence of a strong Middle Eastern signal on Madagascar could be due to sampling bias, since Arab or Persian traders are known to have established posts on the Northwest coast of Madagascar, whereas only populations from the centre and South of Madagascar have been studied to date.[5,8] The low frequencies of E-M293 and A-M91, on both the Comoros and Madagascar, contrasts with the high frequency found in inland populations from Tanzania and Kenya,[13,38] and could be characteristics of a genetic profile specific to sub-Saharan coastal East Africa.

The SEA haplogroups, shared with Madagascar, on the Comoros, are O1a-M50 for the Y chromosome and M7c1c, F3b and B4a1a1-PM for the mitochondria. Consistent with their transit West across the Indian Ocean to the East African coast, O1a-M50, M7c1c and F3b are linked to maritime colonisation within island SEA. In contrast, B4a1a1-PM has not been found in island SEA further West than Southeast Borneo (1%) and has expanded mainly East into Polynesia but also West to Madagascar where it predominates.[25,52] There are nevertheless several indicators that the Comoros' history of gene flow from SEA is distinct from Madagascar's: the absence of Y HgO2-M95 and the very low frequency of B4a1a1-PM, on the Comoros, the higher frequency of F3b (Comoros 8%; Madagascar 3.7%), the dissimilarity of M* HVS-I sequences and the low affinity between the Comoros' O1a-M50 chromosomes and those of Madagascar. The fact that O1a-M50, M7c1c, F3b and B4a1a1-PM have not been found at sites around the Indian Ocean,[26,27,39,45,56] outside SEA and now East Africa, is consistent with a colonising migration from SEA to East Africa directly across the Indian Ocean.

The Comoros population represents an exceptionally diverse genetic mosaic created by the complex process of human settlement around the Indian Ocean. The genealogy of our large sample is well-documented and will provide a solid base from which to explore human diversity in Madagascar and coastal sub-Saharan East Africa.

Supplemental information is available at the *European Journal of Human Genetics* website.

Acknowledgements

We express our sincere gratitude to the people of the Comoros who participated in this study, and the medical personnel of the Comoros and the World Health Organisation for sample collection, and in particular Dr Said Ahamada Fazul, Dr Islam Abdallah, Mr Ismail Msaidié, Mrs Wardat Abdoulkarim, Mme Oumrati Haribou, Mr A Djazza and Mr A.I Abdallah. We thank the EFS (French Blood Transfusion Service) Alpes-méditerranée for their support. We further thank Professor Nicolas Levy, Dr Catherine Badens, Caroline Lacoste, Julie di Cristofaro, Marie-Claude Bonino Fabien Ciné and Nathalie Eudes for help with genotyping.

A.D. was supported by a MENRT post-graduate studentship from the French Government and a Fyssen post-doctoral fellowship. S.M. was supported by a studentship "Coopération France-Comores" given jointly from the French and the Comoros governments. This work was supported by the CNRS grant “OHLL – Origines de l’Homme, des langues et du language” and core funding from Inserm to MJM.

Conflict of interest. The authors declare no conflict of interest.

References

1 Beaujard P: The Indian Ocean in Eurasian and African World-Systems Before the Sixteenth Century. *J World Hist* 2005; **16**: 411-465.

2 Chaudhuri KN: Trade and Civilisation in the Indian Ocean: An Economic History from the Rise of Islam to 1750. Cambridge, Cambridge University Press, 1985.

3 Allibert C: Le mot " KOMR" dans l'océan indien. *Etudes Océan Indien* 2001; **31**: 13-33.

4 Blench R: The ethnographic evidence for long-distance contacts between Oceania and East Africa.; in: Reade J (ed): The Indian Ocean in Antiquity. London / New York, Kegan Paul / British Museum, 1994, pp 461-470.

5 Hurles ME, Sykes BC, Jobling MA, Forster P: The dual origin of the malagasy in island southeast Asia and East Africa: evidence from maternal and paternal lineages. *Am J Hum Genet* 2005; **76**: 894-901.

6 Regueiro M, Mirabal S, Lacau H, Caeiro JL, Garcia-Bertrand RL, Herrera RJ: Austronesian genetic signature in East African Madagascar and Polynesia. *J Hum Genet* 2008; **53**: 106-120.

7 Soodyall H, Jenkins T, Stoneking M: 'Polynesian' mtDNA in the Malagasy. *Nat Genet* 1995; **10**: 377-378.

8 Tofanelli S, Bertoncini S, Castri L *et al*: On the origins and admixture of Malagasy: new evidence from high resolution analyses of paternal and maternal lineages. *Mol Biol Evol* 2009.

9 Nurse D, Hinnebusch TJ: Swahili and Sabaki: A Linguistic History. Berkeley and Los Angeles, University of California Press, 1993.

10 Allibert C: Austronesian Migration and the Establisment of the Malagasy Civilization: Contrasted Readings in Linguistics, Archaeology, Genetics and Cultural Anthropology. *Diogenes* 2008; **55**: 7-16.

11 Allibert C, Vérin P: The Early Pre-Islamic History of the Comores Islands: Links with Madagascar and Africa; in: Reade J (ed): The Indian Ocean in Antiquity. London, New York, Kegan Paul International, British Museum, 1996, pp 461-470.

12 Allibert C: La chronique d'Anjouan par Said Ahmed Zaki; in: Allibert C (ed): Anjouan dans l'histoire. Paris, CEROI-INALCO, 2000, vol 29, pp 9-92.

13 Luis JR, Rowold DJ, Regueiro M *et al*: The Levant versus the Horn of Africa: Evidence for Bidirectional Corridors of Human Migrations. *Am. J. Hum. Genet.* 2004; **74**: 532–544.

14 Semino O, Magri C, Benuzzi G *et al*: Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. *Am J Hum Genet* 2004; **74**: 1023-1034.

15 Spurdle AB, Jenkins T: The origins of the Lemba "Black Jews" of southern Africa: evidence from p12F2 and other Y-chromosome markers. *Am J Hum Genet* 1996; **59**: 1126-1133.

16 Badens C, Martinez di Montemuros F, Thuret I *et al*: Molecular basis of haemoglobinopathies and G6PD deficiency in the Comorian population. *Hematol J* 2000; **1**: 264-268.

17 Chiaroni J, Touinssi M, Mazet M, De Micco P, Ferrera V: Adsorption of autoantibodies in the presence of LISS to detect alloantibodies underlying warm autoantibodies. *Transfusion* 2003; **43**: 651-655.

18 Gibert M, Touinssi M, Reviron D, Mercier P, Boetsch G, Chiaroni J: HLA-DRB1 frequencies of the Comorian population and their genetic affinities with Sub-Saharan African and Indian Oceanian populations. *Ann Hum Biol* 2006; **33**: 265-278.

19 Miller SA, Dykes DD, Polesky HF: A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* 1988; **16**: 1215.

20 Behar DM, Villems R, Soodyall H *et al*: The dawn of human matrilineal diversity. *Am J Hum Genet* 2008; **82**: 1130-1140.

21 Excoffier L, G. L, Schneider S: Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* 2005; **1**: 47-50.

22 Reynolds J, Weir BS, Cockerham CC: Estimation of the Coancestry Coefficient: Basis for a Short-Term Genetic Distance. *Genetics* 1983; **105**: 767-779.

23 Slatkin M: A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 1995; **139**: 457-462.

24 Dupanloup I, Bertorelle G: Inferring admixture proportions from molecular data: extension to any number of parental populations. *Mol Biol Evol* 2001; **18**: 672-675.

25 Hill C, Soares P, Mormina M *et al*: A mitochondrial stratigraphy for island southeast Asia. *Am J Hum Genet* 2007; **80**: 29-43.

26 Metspalu M, Kivisild T, Metspalu E *et al*: Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. *BMC Genet* 2004; **5**: 26.

27 Regueiro M, Cadenas AM, Gayden T, Underhill PA, Herrera RJ: Iran: tricontinental nexus for Y-chromosome driven migration. *Hum Hered* 2006; **61**: 132-143.

28 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.

29 Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF: New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res* 2008; **18**: 830-838.

30 Kayser M, Brauer S, Cordaux R *et al*: Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. *Mol Biol Evol* 2006; **23**: 2234-2244.

31 Karafet TM, Lansing JS, Redd AJ *et al*: Balinese Y-chromosome perspective on the peopling of Indonesia: genetic contributions from pre-neolithic hunter-gatherers, Austronesian farmers, and Indian traders. *Hum Biol* 2005; **77**: 93-114.

32 Cerny V, Pereira L, Kujanova M *et al*: Out of Arabia-The settlement of Island Soqotra as revealed by mitochondrial and Y chromosome genetic diversity. *Am J Phys Anthropol* 2008.

33 Cruciani F, Santolamazza P, Shen P *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.

34 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.

35 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.

36 Underhill PA, Shen P, Lin AA *et al*: Y chromosome sequence variation and the history of human populations. *Nat Genet* 2000; **26**: 358-361.

37 Sims LM, Garvey D, Ballantyne J: Sub-populations within the major European and African derived haplogroups R1b3 and E3a are differentiated by previously phylogenetically undefined Y-SNPs. *Hum Mutat* 2007; **28**: 97.

38 Henn BM, Gignoux C, Lin AA *et al*: Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. *Proc Natl Acad Sci U S A* 2008; **105**: 10693-10698.

39 Cadenas AM, Zhivotovsky LA, Cavalli-Sforza LL, Underhill PA, Herrera RJ: Y-chromosome diversity characterizes the Gulf of Oman. *Eur J Hum Genet* 2008; **16**: 374-386.

40 Cinnioglu C, King R, Kivisild T *et al*: Excavating Y-chromosome haplotype strata in Anatolia. *Hum Genet* 2004; **114**: 127-148.

41 Arredi B, Poloni ES, Paracchini S *et al*: A predominantly neolithic origin for Y-chromosomal DNA variation in North Africa. *Am J Hum Genet* 2004; **75**: 338-345.

42 Hassan HY, Underhill PA, Cavalli-Sforza LL, Ibrahim ME: Y-chromosome variation among Sudanese: restricted gene flow, concordance with language, geography, and history. *Am J Phys Anthropol* 2008; **137**: 316-323.

43 Tofanelli S, Ferri G, Bulayeva K *et al*: J1-M267 Y lineage marks climate-driven pre-historical human displacements. *Eur J Hum Genet* 2009.

44 Cruciani F, La Fratta R, Trombetta B *et al*: Tracing past human male movements in northern/eastern Africa and western Eurasia: new clues from Y-chromosomal haplogroups E-M78 and J-M12. *Mol Biol Evol* 2007; **24**: 1300-1311.

45 Sengupta S, Zhivotovsky LA, King R *et al*: Polarity and temporality of high-resolution Y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. *Am J Hum Genet* 2006; **78**: 202-221.

46 Capelli C, Onofri V, Brisighelli F *et al*: Moors and Saracens in Europe: estimating the medieval North African male legacy in southern Europe. *Eur J Hum Genet* 2009; **17**: 848-852.

47 Walsh B: Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals. *Genetics* 2001; **158**: 897-912.

48 Alshamali F, Pereira L, Budowle B, Poloni ES, Currat M: Local population structure in Arabian Peninsula revealed by Y-STR diversity. *Hum Hered* 2009; **68**: 45-54.

49 Roewer L, Willuweit S, Stoneking M, Nasidze I: A Y-STR database of Iranian and Azerbaijanian minority populations. *Forensic Sci Int Genet* 2009; **4**: e53-55.

50 Capelli C, Wilson JF, Richards M *et al*: A predominantly indigenous paternal heritage for the Austronesian-speaking peoples of insular Southeast Asia and Oceania. *Am J Hum Genet* 2001; **68**: 432-443.

51 Kayser M, Brauer S, Weiss G, Schiefenhovel W, Underhill PA, Stoneking M: Independent histories of human Y chromosomes from Melanesia and Australia. *Am J Hum Genet* 2001; **68**: 173-190.

52 Kayser M, Choi Y, van Oven M *et al*: The impact of the Austronesian expansion: evidence from mtDNA and Y chromosome diversity in the Admiralty Islands of Melanesia. *Mol Biol Evol* 2008; **25**: 1362-1374.

53 Li H, Wen B, Chen SJ *et al*: Paternal genetic affinity between Western Austronesians and Daic populations. *BMC Evol Biol* 2008; **8**: 146.

54 Gonder MK, Mortensen HM, Reed FA, de Sousa A, Tishkoff SA: Whole-mtDNA genome sequence analysis of ancient African lineages. *Mol Biol Evol* 2007; **24**: 757-768.

55 Gray RD, Jordan FM: Language trees support the express-train sequence of Austronesian expansion. *Nature* 2000; **405**: 1052-1055.

56 Abu-Amero KK, Gonzalez AM, Larruga JM, Bosley TM, Cabrera VM: Eurasian and African mitochondrial DNA influences in the Saudi Arabian population. *BMC Evol Biol* 2007; **7**: 32.

57 Richards M, Rengo C, Cruciani F *et al*: Extensive female-mediated gene flow from sub-Saharan Africa into near eastern Arab populations. *Am J Hum Genet* 2003; **72**: 1058-1064.

58 Kivisild T, Reidla M, Metspalu E *et al*: Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. *Am J Hum Genet* 2004; **75**: 752-770.

59 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.

60 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.

61 Chang YM, Perumal R, Keat PY, Kuehn DL: Haplotype diversity of 16 Y-chromosomal STRs in three main ethnic populations (Malays, Chinese and Indians) in Malaysia. *Forensic Sci Int* 2007; **167**: 70-76.

62 Wu FC, Ho CW, Pu CE *et al*: Y-chromosomal STRs haplotypes in the Taiwanese Paiwan population. *Int J Legal Med* 2010; DOI: 10.1007/s00414-009-0416-x.

63 Chang YM, Swaran Y, Phoon YK *et al*: Haplotype diversity of 17 Y-chromosomal STRs in three native Sarawak populations (Iban, Bidayuh and Melanau) in East Malaysia. *Forensic Sci Int Genet* 2009; **3**: e77-80.

64 Alam S, Ali ME, Ferdous A, Hossain T, Hasan MM, Akhteruzzaman S: Haplotype diversity of 17 Y-chromosomal STR loci in the Bangladeshi population. *Forensic Sci Int Genet* 2009; **4**: e59-60.

65 Turchi C, Buscemi L, Giacchino E *et al*: Polymorphisms of mtDNA control region in Tunisian and Moroccan populations: an enrichment of forensic mtDNA databases with Northern Africa data. *Forensic Sci Int Genet* 2009; **3**: 166-172.

66 Sanchez JJ, Hallenberg C, Borsting C, Hernandez A, Morling N: High frequencies of Y chromosome lineages characterized by E3b1, DYS19-11, DYS392-12 in Somali males. *Eur J Hum Genet* 2005; **13**: 856-866.

Titles and Legends to Figures

Figure 1. Frequencies (%) and numbers (n) of Y haplogroups in the Comoros population sample. Haplogroup names follow the 2008 nomenclature.[29] Branches are labelled with the binary markers tested. Numbers without a letter represent "M" prefixed Y markers (e.g; 50=M50).[60] Putative geographic origin is indicated for each haplogroup: Af – sub-Saharan Africa, WSA – West and Southwest Asia, SEA – Southeast Asia and ? – uncertain. Frequencies of less than 5% have been rounded up or down to the nearest unit.

Figure 2. Multidimensional scaling (MDS) analysis plot of genetic distance (Rst) calculated from the incidence of alleles at eight Y STR loci (DYS19, 389AB, 389CD, 390, 391, 392, 393, 439). The analysis was performed with subsets of the Comoros sample, which were created on the basis of putative haplogroup origin. a. Middle East - haplogroups E-M123, E-V22, F, G, J, L, Q and R. b. Southeast Asian –haplogroups O, C* and K*. The populations represented are the Comoros (COM), this study, Madagascar (MAD),[8] Oman (OMA),[13,48] Turkey (TUR),[40] North Pakistan (N-PAK), South Pakistan (S-PAK), North India (N-IND), South India (S-IND),[45] Yemen (YEM), United Arab Emirates (UAE), Saudi Arabia (SAU),[48], North Iran (N-IR),[49] South Iran (S-IR),[48,49] Malaysia (MAL),[61] Taiwan (Paiwan) (TAI),[62] West Borneo (East Malaysia) (W-BOR)[63] and Bangladesh (BAN).[64]

Figure 3. Frequencies (%) and numbers (n) of mitochondrial haplogroups in the Comoros population sample. Numbers on branches refer to the position of polymorphisms in the CRS (Cambridge reference sequence). HVS-I sequence was not determined for L0, L1, L2 or L3'4(xMN). The HVS-I SNPs are shown for M and N haplogroups, only where they provide further definition than the coding SNPs. Putative

geographic origin is indicated for each haplogroup: Af – sub-Saharan Africa, SEA – Southeast Asia and ? – uncertain.

Figure 4. Multidimensional scaling (MDS) analysis plot of genetic distance (Fst) calculated from mitochondrial haplogroup frequencies. M* and R* were excluded from these analyses. a. Africa, SEA and Iran – all Comoros haplogroups except M* and R*. b. and c. MDS performed with subsets of the Comoros sample, defined on the basis of putative haplogroup origin. b. Africa – Comoros haplogroups L. c. SEA – Comoros and Madagascar haplogroups B4a, B4a1a1-PM, F3b, M7c1c and R9. The populations are Comoros (COM), this study, Madagascar (MAD),[8] Central Africa (AFC),[54] Iran (IRA), Mozambique (MOZ), Kenya (KEN),[56] Ethopia (ETH),[58] Tunisia (TUN), Algeria (ALG), Morocco (MOR), Mauritania (MAU),[65] Taiwan (TAI), Philippines (PHI), Malaysia (MAL), Borneo (BOR), Sumatra (SUM), Bali (BAL), Java (JAV).[25]

Supplemental Figure 1. Distribution of the 26 sites sampled from the three islands of the Comoros Republic, A - Grand Comore (*Ngazidja*), B - Anjouan (*Ndzuani*), C - Mohéli (*Mwali*). We did not sample from the French Island, D – Mayotte (*Maore*).

Supplemental Figure 2. The spread of Western Eurasian Y chromosomes and trade routes around the Western rim of the Indian Ocean. **A:** Incidence of Western Eurasian Y haplogroups E-V22, E-M123, G, I, J, L Q and R. Pie charts show the incidence of Northern haplogroups in colour with other haplogroups in white, except 1b which represents the relative incidence of the Northern haplogroups on the Comoros, expanded to 66% of the pie chart to facilitate comparison with Iran and Turkey. Populations: 1a), 1b) Comoros – this study, 2a) Tanzania, 2b) Kenya,[13] 3) Somalia,[66] 4) Yemen,[39] 5) Oman, 6) Egypt (*G-M201),[13] 7) Turkey (*L(xM349),[40] 8) Pakistan, 9) India[45], 10) South Iran[27] and 11) Madagascar[8] and 12) United Arab Emirates.[39] The absence of these haplogroups between the Middle East and the Comoros emphasises that their migrations South to the Comoros were by maritime routes. Present-day Iran is the most likely area of origin for the Western Asian Y chromosomes on the Comoros. **B:** The development of trade routes and towns, based on archaeological and historical evidence.[1] It can be seen that some of the first towns that grew up around the earliest trade routes are situated in present-day Iran.

Supplemental Figure 3. Mitochondrial and Y gene flow from Southeast Asia across the Indian Ocean to the East coast of Africa. The incidence of mitochondrial haplogroups common to the Comoros and Southeast Asia are represented by pie charts. The frequency of Y haplogroup O1a2 (defined by M50 or its putative phylogenetic

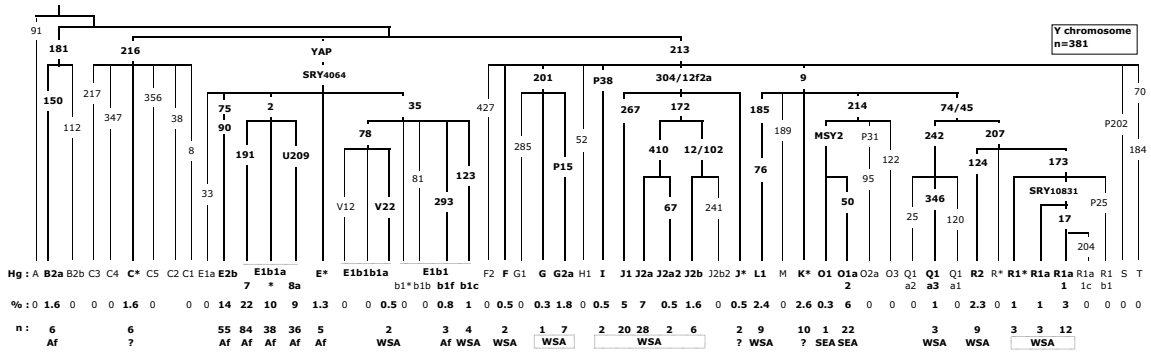
equivalent M110) is shown boxed. Populations shown are 1) Comoros –this study, 2) Madagascar,[5,8] and 3) Sumatra, 4) Borneo, 5) Sulawesi, 6) Malaysia, 7) Philippines, 8) Taiwan.[5,25,52,53]

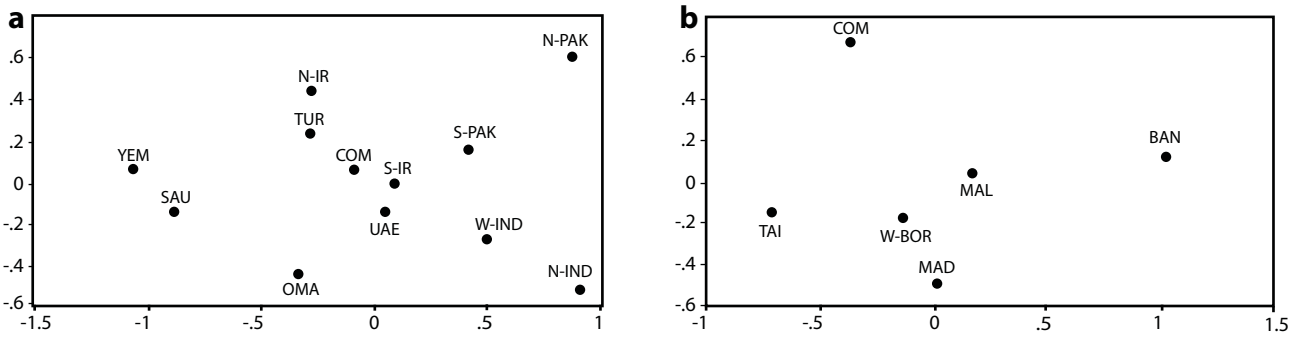
Supplemental Table 1. Y chromosome SNP markers.

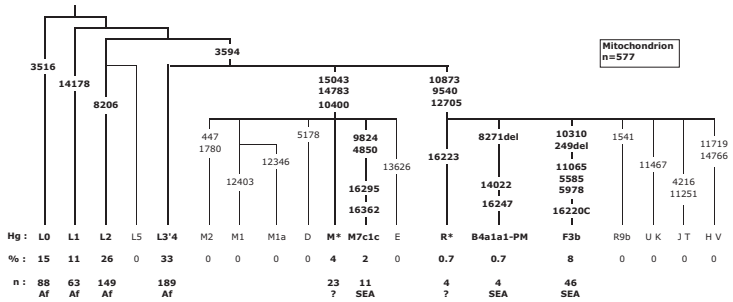
Supplemental Table 2. Y chromosome haplogroups and STR haplotypes.

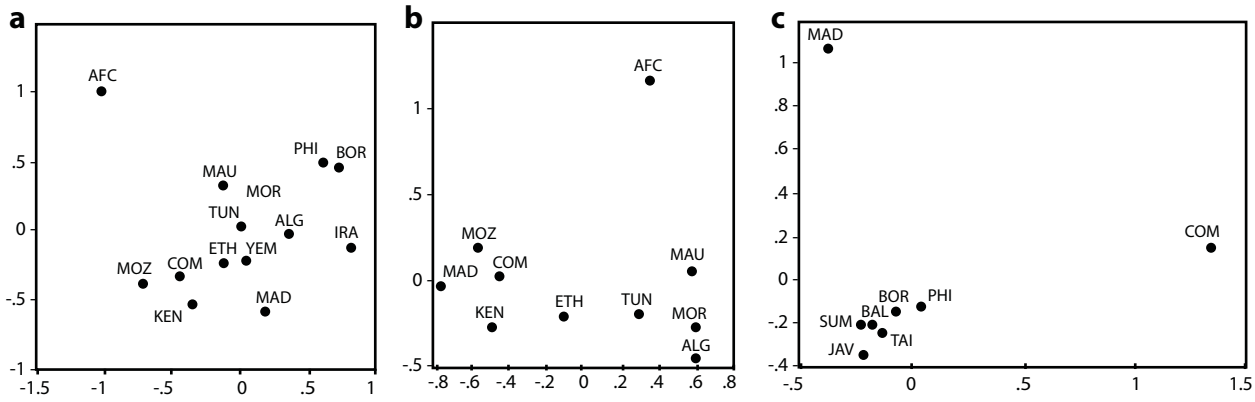
Supplemental Table 3. Mitochondrial SNP markers.

Supplemental Table 4. Mitochondrial haplogroups and HVS-I sequence.


Y chromosome
n=381
Hg : A B2a B2b C3 C4 C* C5 C2 C1 E1a E2b E1b1a* E* E1b1b1a E1b1b1 F2 F G1 G G2a H1 I J1 J2a J2a2 J2b J2b2 J* L1 M K* O1 O1a O2a O3 Q1 a2 Q1 a3 Q1 a1 R2 R* R1* R1a R1a 1 R1a 1c R1 b1 S T
% : 0 1.6 0 0 0 1.6 0 0 0 14 22 10 9 1.3 0 0 0.5 0 0 0.8 1 0 0.5 0.3 1.8 0 0.5 5 7 0.5 1.6 0 0.5 2.4 0 2.6 0.3 6 0 0 0 1 0 2.3 0 1 1 3 0 0 0 0
n : 6 Af 6 ? 55 Af 84 Af 38 Af 36 Af 5 Af 2 WSA 3 Af 4 WSA 2 WSA 1 7 WSA 2 20 28 2 6 WSA 2 9 ? WSA 10 1 22 ? SEA SEA 3 WSA 9 WSA 3 3 12 WSA
91 181 216 YAP SRY4064 213 150 112 217 347 356 38 8 33 75 90 191 2 U209 78 V12 V22 35 81 293 123 427 285 201 P15 52 P38 267 304/12f2a 410 67 172 12/102 241 185 76 189 9 214 MSY2 50 P31 95 122 74/45 242 346 25 120 207 124 173 SRY10831 17 204 P25 P202 70 184


a
N-PAK
N-IR
TUR
S-PAK
YEM
COM
S-IR
SAU
UAE
W-IND
N-IND
OMA
.6
.4
.2
0
-.2
-.4
-.6
-1.5
-1
-.5
0
.5
1
b
COM
BAN
MAL
TAI
W-BOR
MAD
.6
.4
.2
0
-.2
-.4
-.6
-1
-.5
0
.5
1
1.5


Mitochondrion
n=577
3516
14178
8206
3594
15043
14783
10400
10873
9540
12705
447
1780
12403
12346
5178
9824
4850
16295
16362
13626
16223
8271del
14022
16247
10310
249del
11065
5585
5978
16220C
1541
11467
4216
11251
11719
14766
Hg :
L0
L1
L2
L5
L3'4
M2
M1
M1a
D
M*
M7c1c
E
R*
B4a1a1-PM
F3b
R9b
U K
J T
H V
% :
15
11
26
0
33
0
0
0
0
4
2
0
0.7
0.7
8
0
0
0
0
n :
88
Af
63
Af
149
Af
189
Af
23
?
11
SEA
4
?
4
SEA
46
SEA


a
AFC
PHI
BOR
MAU
MOR
TUN
ALG
IRA
ETH
YEM
MOZ
COM
MAD
KEN
1
.5
0
-.5
-1
-1.5
-1
-.5
0
.5
1
b
AFC
MOZ
MAD
COM
MAU
ETH
TUN
KEN
MOR
ALG
1
.5
0
-.5
-.8
-.6
-.4
-.2
0
.2
.4
.6
.8
c
MAD
COM
BOR
PHI
BAL
SUM
TAI
JAV
1
.8
.6
.4
.2
0
-.2
-.4
-.5
0
.5
1
1.5

Haplogroup	Mutation	n	Variance
B2a	M150	6	1.01
C*	M216	6	0.44
E*	SRY_{4064}	5	1.03
E1b1a*	M2	15	0.46
E1b1a7	M191	19	1.13
E1b1a8a	U209	36	0.66
E2	M75	14	0.75
G2a	P15	7	0.10
J1	M267	20	0.54
J2a	M410	28	0.60
J2b	M12	6	0.57
K*	M9	10	1.16
L1	M76	9	1.05
O1a2	M50	22	0.30
R1a1	M17	12	0.77
R2	M124	9	1.18

Table 1 Variance of the principal Y haplogroups (n ≥5) on the Comoros based on 15 Y-microsatellite loci. The markers DYS385a and b were not used in the calculation. The variance is calculated from STR data for all individuals of each haplogroup (n), except E1b1a7 (19 of 84 men) and E1b1a* (15 of 38 men).