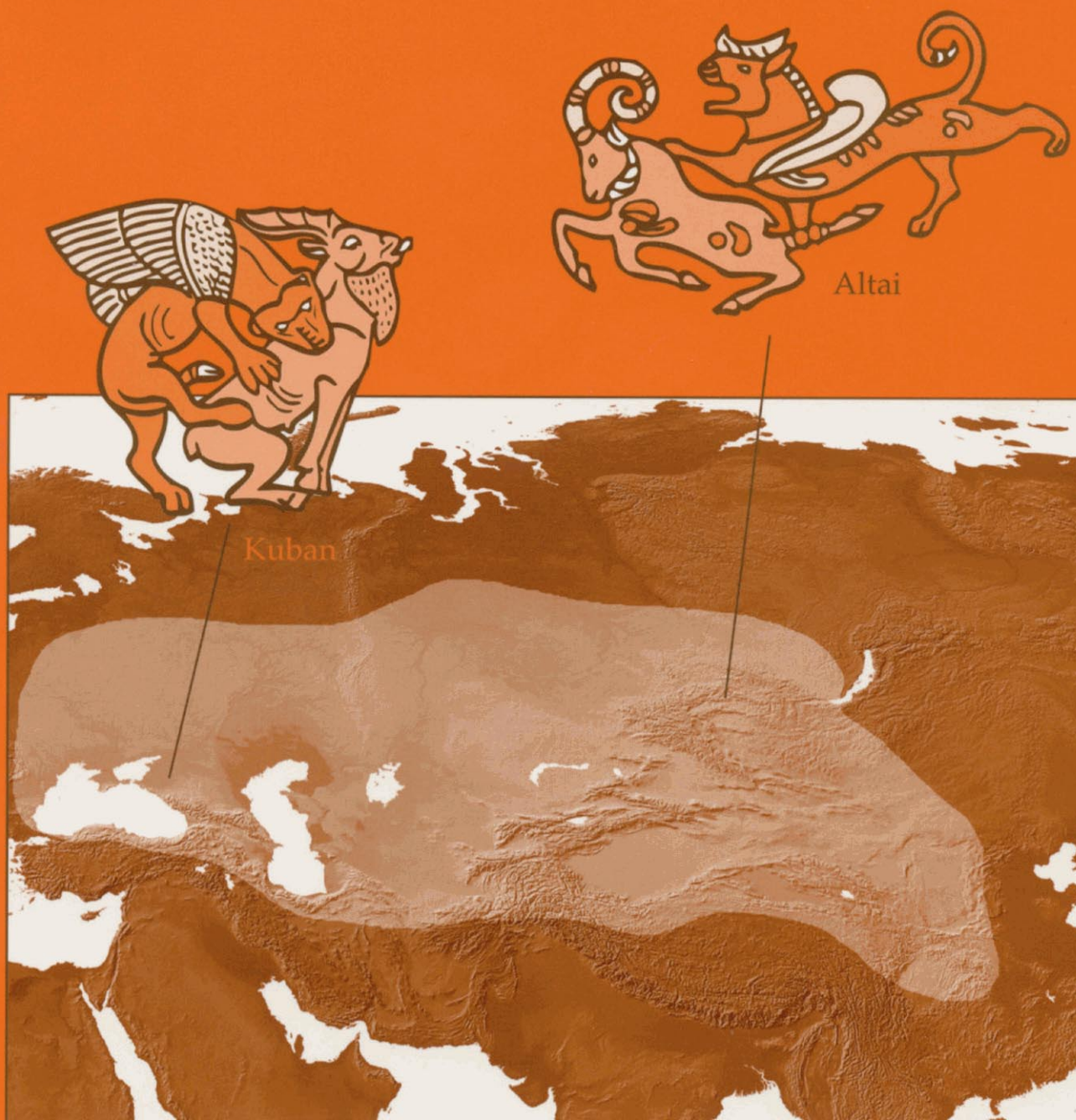
McDONALD INSTITUTE MONOGRAPHS

Ancient interactions: east and west in Eurasia

Edited by Katie Boyle, Colin Renfrew & Marsha Levine



Chapter 20

Analysis of Y-chromosome Variation in Modern Populations at the European–Asian Border

Patrizia Malaspina, Andrey I. Kozlov, Fulvio Cruciani, Piero Santolamazza, Nejat Akar, Dimiter Kovatchev, Marina G. Kerimova, Juri Parik, Richard Villems, Rosana Scozzari & Andrea Novelletto

Population genetic data have long been sought as a relevant resource for the understanding of population affinities and divergence. In particular, the present-day genetic diversity observed among human populations is the result of their long-term isolation in the past, when population size was much smaller owing to ecological constraints. Under these conditions, reduced gene flow caused by geographical barriers as well as cultural, linguistic and religious differences, yielded a process of genetic divergence among populations. This process is regarded as having occurred over long periods of prehistoric time as compared to recorded history, and to have involved all portions of the human genome whose variation does not strongly influence the ability of the organism to survive in a specific environment (neutral polymorphism).

It is only after the advent of molecular genetics techniques that genetic variants (alleles) at neutral loci could be fully characterized. This allowed the reconstruction of phylogenies, that is the order of appearance of lineages characterized by new mutations (sequence changes or rearrangements) introduced on the pre-existing DNA sequences. In this context, DNA markers of the genetic diversity of the human Y chromosome are gaining an ever-increasing value in understanding human microevolution. In fact, the hemizygous state (each male carries a single Y chromosome as opposed to a double dose of the 22 autosomal chromosomes) and uniparental inheritance (i.e. each male inherits his Y chromosome from his father only) render the non-recombinant portion of this chromosome (NRPY) a unique system to describe the progressive accumulation of mutations and to make inferences based on the increase of diversity over time. In previous works we have shown that the combined use of binary (such as Single Nucleotide Substitutions and Alu insertions, which only have two possible alleles) and microsatellite (which have an array of different length alleles) markers greatly increase the possibility of detecting different Y-chromosomal lineages in the gene pools of extant populations. In fact, binary polymorphisms can be considered the result of rare or even unique events. In other words, chromosomes carrying the same variant (allele) at a given locus (a precise position along the chromosome) can be considered related by descent. A combination of these alleles (haplotype) can therefore be used to reconstruct a worldwide phylogeny of the Y chromosome. Complementary to this, alleles for the fast-evolving microsatellites can not be related by descent worldwide but increase the possibility of defining Y-specific phylogenies on smaller geographical and time scales.

Materials and methods

The subjects

We studied males corresponding to 15 locations in North and Northeastern Europe and West Asia (Table 20.1). Each group was well characterized according to his parental origin and spoken language. DNA was prepared by standard techniques from either fresh venous blood or dried blood absorbed on filter paper.

The markers

Four binary polymorphisms and two complex dinucleotide microsatellite systems were studied. The

presence / absence of the YAP element, of the alphoid HindIII and of the A/G transition at DYS257 were assayed by PCR as already described (Hammer & Horai 1995; Santos *et al.* 1995; Hammer *et al.* 1998). The A/G base substitution at position 10831 of the SRY sequence (SRY_{10831} or SRY-1532) (Whitfield *et al.* 1995) was assayed with two previously-described independent methods (Santos *et al.* 1999; Scozzari *et al.* 1999). These markers allowed us to distinguish haplotypes 1B, 1C and ID and another group which includes haplotypes 3G, 3A, 4 and 5 (all YAP+ chromosomes) as defined by Hammer and colleagues (1998).

YCAII and DYS413 polymorphic systems were detected according to Mathias *et al.* (1994) and Malaspina *et al.* (1997). These systems consist of two Y-specific loci each, both containing a (CA)n microsatellite, that are co-amplified during the corresponding Polymerase Chain Reactions (PCR).

Statistical analyses

One-step networks of adjacent microsatellite haplotypes on chromosomes with the same allelic states for binary polymorphisms were constructed as described (Malaspina *et al.* 1998). Briefly, each network groups only 'adjacent' haplotypes, i.e. haplotypes that differ for the insertion or deletion of a single CA unit at a single locus. The method attempts to infer microsatellite phylogeny based on the observation that one-step mutations are the predominant type of mutation in microsatellites.

The chromosomes reported here were assigned or adjoined to one of the one-step networks constructed with an initial set of 908 Eurasian chromosomes (Malaspina *et al.* 1998), then extended to 1801 chromosomes (Malaspina *et al.* 2000). This led to four large (1.1, 1.2, 2.1, 3.1) and several small networks, plus a minority of unclassified haplotypes. Network 3.1 reported in Malaspina *et al.* (1998) has now been subdivided into 3.1G and 3.1A according to SRY_{10831} typing. Networks 1.1 and 1.2 group different subsets of chromosomes classified as 1B based on SNS; network 2.1 groups YAP+ chromosomes, mostly of haplotype 4 (Malaspina unpublished); networks 3.1G and 3.1A consist of chromosomes classified as 1C and ID, respectively.

The overall network frequencies in the 15 populations were used to compute F statistics, a measure of divergence between populations and groups of populations obtained by partitioning the total genetic variance of the sample into the components determined by the hierarchical grouping of subjects (Table 20.2). All F statistics were obtained by using the Arlequin package v. 1.1 (Schneider *et al.* 1997).

Results

The frequency of binary haplotypes varied significantly between the 15 populations ($\chi^2 = 32$, 42 d.f., $p < 10^{-5}$). In particular, all Continental Turkish populations showed haplotype 1B frequencies above 69 per cent while this haplotype reached only 35 per cent among the Talysh. Turkish Cypriots had the highest frequencies of YAP+ haplotypes (32 per cent), followed by the Bulgarians (18 per cent). Haplotype 1C reached a frequency of 50 per cent in the Talysh, as compared to <17 per cent in all other populations (9 per cent in the entire data set). Haplotype ID had

Table 20.1. *The population samples studied and their network frequencies (in percentages).*

Region	Population	Linguistic family	Sample size *(n)*	Network									
				1.1	1.2	2.1	3.1G	3.1A	1.4	1.7	1.5	3.6A	unc.
North Europe	Estonian	Uralic	74	54.1	1.4	4.1	2.7	36.5					1.4
	Komi-Permiak	Uralic	42	64.3	4.8	2.4	4.8	23.8					
	Russian (Perm)	I.E.	37	35.1	2.7		13.5	43.2					5.4
	Moldovan Erzya	Uralic	46	41.3	4.3	2.2	6.5	39.1				6.5	
	Moldovan Moksi	Uralic	46	65.2	6.5	2.2	2.2	21.7					2.2
	Estonia Russian	I.E.	26	57.7	3.8		11.5	26.9					
Central Europe	Ukrainian	I.E.	6	33.3		16.7		50.0					
	Bulgarian	I.E.	34	47.1	11.8	14.7	5.9	14.7					5.9
West Asia	Northeast Turkish	Altaic	11	54.5	9.1		9.1	18.2		9.1			
	Central Anatolian	Altaic	18	33.3	22.2	5.6	5.6	11.1		5.6	5.6		11.1
	Southwest Turkish	Altaic	29	37.9	31.0	10.3	3.4	10.3	3.4				3.4
	Southeast Turkish	Altaic	13	46.2	23.1		7.7	15.4					7.7
	Turkish Cypriots	Altaic	22	18.2	27.3	22.7	4.5	13.6		4.5			9.1
	Talysh	I.E.	20	15.0	20.0	5.0	45.0	10.0					5.0
	Azeri	Altaic	24	33.3	20.8	8.3	16.7	12.5	4.2	4.2			

frequencies >20 per cent in all North and North-Eastern European populations, dropping both in West Asia and in Bulgaria.

The identification of one-step networks based on microsatellite data allowed an even higher resolution (Table 20.1). Overall, 435 out of 448 chromosomes could be assigned to one of 9 previously-detected one-step networks. In particular, network 1.2 accounted for (>20 per cent of all chromosomes in four out of five Turkish populations, in the Azeri and in the Talysh. Network 2.1 grouped 24 out of 28 YAP+ chromosomes. Network 3.1G grouped 36 of the 40 1C chromosomes. This network reached a frequency of 45 per cent in the Talysh, and was mainly represented by haplotype 23-19-23-23 (7 out of 9), identical to that highly prevalent in chromosomes from Western Europe. Network 3.1A grouped 113 out of the 116 chromosomes ID. The single microsatellite haplotype 23-19-22-22, already observed as the major haplotype of this network, was also highly prevalent in the populations reported here. The remaining 3 haplotypes ID were identical and were only found in the Moldovan Erzya.

We performed the analysis of *F* statistics based on network frequencies (Table 20.2). The results show an overall *Fst* of 0.07 ($p < 10^{-5}$). Table 20.3 reports *Fst* values obtained for all pairwise comparisons between populations. One can observe that the Talysh and the Turkish Cypriots produced values >0.10 in 10 out of 14 and 6 out of 14 comparisons, respectively. This excess of high values can be attributed to the outlying frequencies of networks 3.1G and 2.1 in the two populations, respectively (Table 20.1). On the contrary, *Fst* values between Continental Turkish populations are all very low.

The role of linguistics

Table 20.2 reports the *F* values obtained after grouping the populations according to their linguistic family. The percentage of variation among groups (linguistic families) is larger than within groups, with a highly significant value of *Fst*. As for comparisons within the same linguistic family, significant heterogeneity is observed only among Indo-European-speakers. Pairwise *Fst* analysis (Table 20.3) shows very low values for all comparisons between speakers of Altaic Ianguages. The highest *Fst* values are found in all comparisons between the Talysh- and Uralic-speakers. However, these latter comparisons correspond to geographic distances >1600 km in all cases. It is not surprising that an appreciable genetic divergence is accumulated between populations so far apart, more so in the presence of such a relevant barrier as the Caucasus.

Three pairwise comparisons allowed us to evaluate the genetic divergence of sympatric or quasi-sympatric populations speaking languages from different families. Only in the case of Komi-Permiaks vs. Perm Russians was a significant *Fst* value obtained. In particular, the genetic similarity between Estonia Russians and Estonians, as well as other Uralic-speakers, can be attributed to the acquisition of the Russian language by the former.

Discussion

The data reported here are useful in interpreting the composition of the present-day populations of Eastern Europe and Western Asia, especially if one considers the relative antiquity of the different Y-chromosomal lineages here detected. Further inferences can be drawn by considering the degree of divergence of populations living nearby and the linguistic families to which they belong.

Network 1.2 is the most recent among the lineages detected by the markers used in the populations here considered. Its antiquity has been estimated at approx. 300 generations ago from the extended data set (Malaspina *et al.* 1998). The home-range of this network appears to span from Turkey eastward to-

Table 20.2. *Structuring of the populations here studied after grouping according to linguistic affiliation.*

	Variance components (d.f.)	Per cent		
All populations				
Among groups	0.015 (2)	4.19	*Fct* = 0.042	$p = .007$
Among populations / within groups	0.009 (12)	2.58	*Fsc* = 0.027	$p < 10^{-5}$
Within populations	0.335 (433)	93.23	*Fst* = 0.068	$p < 10^{-5}$
Indo-European-speakers				
Among populations	0.029 (4)	7.60	*Fst* = 0.076	$p < 10^{-3}$
Within populations	0.348 (118)	92.40		
Uralic-speakers				
Among populations	0.005 (3)	1.56	*Fst* = 0.015	p = n.s.
Within populations	0.292 (204)	98.44		
Altaic-speakers				
Among populations	-0.005(5)	-1.25	*Fst* = -0.01	p = n.s.
Within populations	0.400 (111)	101.25		

Table 20.3. *Fst obtained in all pairwise comparisons between populations and their significance levels. Populations are reported in the same order as in Table 20.1.*

	Estonian	Komi -P.	Perm R.	M. Erzya	M. Moksi	E. Russ.	Ukrain.	Bulg.	NE Turk.	C. Anat.	SW Turk.	SE Turk.	T. Cyp.	Talysh
Komi-P.	0.007													
Perm R.	0.024	0.078°												
M. Erzya	0.002	0.041	–0.011											
M. Moksi	0.015	–0.021	0.093°	0.053										
E. Russ.	–0.008	–0.022	0.028	0.007	–0.013									
Ukrain.	–0.024	0.069	–0.061	–0.061	0.086	0.021								
Bulg.	0.039°°	0.022	0.060°	0.038°	0.022	0.011	0.014							
NE Turk.	–0.007	–0.037	0.026	0.000	–0.033	–0.051	0.010	–0.032						
C. Anat.	0.090°°	0.082°	0.063°	0.055°	0.079°	0.052°	0.030	–0.011	–0.018					
SW Turk.	0.111°°	0.093°	0.104°°	0.083°°	0.089°	0.074°	0.068	0.003	0.008	–0.030				
SE Turk.	0.041°	0.018	0.042°	0.025	0.015	–0.003	0.030	–0.030	–0.059	–0.051	–0.036			
T. Cypriots	0.165°°	0.179°°	0.111°°	0.109°°	0.180°°	0.141°°	0.041	0.036	0.070	–0.016	–0.000	0.021		
Talysh	0.244°°	0.250°°	0.138°	0.168°°	0.262°°	0.180°°	0.161°	0.126°	0.126	0.067	0.105°	0.089	0.074	
Azeri	0.091°°	0.081°	0.058°	0.052°	0.086°	0.044	0.032	–0.003	–0.017	–0.031	–0.017	–0.034	–0.000	0.030

° $P < 0.05$
°° $P < 0.001$

wards the Caspian Sea and to exclude areas inhabited by Uralic-speakers. Populations speaking very different languages, i.e. Indo-European (the Talysh) and Altaic (the Turkish and the Azeri), show little differences in the incidence of this network. One can hypothesize that the presence of this network in this area is the result of migrations of Altaic-speakers into Turkey in the eleventh century AD and subsequent admixture into the Talysh. This hypothesis is, however, ruled out by the high (20-30 per cent) frequency of this network in Continental Greece and Crete. This would imply a massive contribution of recent Turkish genes to the gene pools of these populations, an unlikely occurrence. Thus it is reasonable to conclude that network 1.2 chromosomes were peculiar to Indo-European-speaking populations which settled in and in the surroundings of the Turkish peninsula prior to the arrival of Altaic-speakers.

Network 3.1A also shows a much stronger association with geography than linguistics. It is the second most recent network and its frequency peaks in the northeastern part of the area here considered, in populations speaking either Indo-European or Uralic languages. Chromosomes carrying the characterizing mutation of this network, SRY_{10831}-A, have been found at high frequencies in Central Asia and Pakistan (Santos *et al.* 1999). Our data then depict an area of intrusion of Asian genes in northeastern Europe, at least up to the Eastern coast of the Baltic Sea. The Altaic-speakers here considered show much lower frequencies of this network. The internal structuring of these groups is virtually null. This is compatible with both a recent splitting of the group ancestral to present-day populations, and a continuous massive gene flow among populations of the group here examined.

Interestingly, the Talysh, an Indo-European-speaking population of the southern coast of the Caspian Sea, show a high incidence of network 3.1G, a feature that establishes a similarity with other Indo-European speakers from Western Europe. This network appears to be much older than networks 1.2 and 3.1A. Santos *et al.* (1999) have placed the origin of the chromosomes related to this type in Central Asia and have postulated an ancient migration westward to explain their high frequency in Western Europe. Present data yield a pre-existing home range of network 3.1G chromosomes extending from West Asia to the Atlantic Sea that, south of the Caucasus, is split by Altaic-speaking newcomers in the region of modern Turkey and Azerbaijan. We recently described (Scozzari *et al.* 1999) populations in Northern Cameroon (Africa) with a high frequency of

haplotype 1C (to which network 3.1G belongs), but distinct microsatellite length alleles. This would testify an ancient migration into Africa from haplotype 1C homeland. The hypothesis that chromosomes 1C in the Talysh are a relict of such a movement from Asia to Africa west of the Caspian Sea, is at odds with the finding of microsatellite haplotypes different from the African ones yet identical to those of Western Europe.

Overall, the data presented here reveal a variety of inputs from Central Asia into the gene pools of West Asian and Northeastern European populations. These flows occurred either north or south of that chain of obstacles as the Black Sea, the Caucasus and the Caspian Sea. However, Y chromosomes travelling with such migrations seem to derive from distinct subsets of the overall diversity that is being revealed by Central Asia.

Acknowledgements

Work supported by grants PRIN MURST 1997, 1999 (RS) and CNR 98.00485.CT04 (AN), 97.00702.PF36 (RS), 97.00712.PF36. Sampling campaigns were partially funded by NATO grant LST.CLG975057 to AIK, MGK and AN.

References

Hammer, M.F. & S. Horai, 1995. Y-chromosomal DNA variation and the peopling of Japan. *American Journal of Human Genetics* 56, 951-62.

Hammer, M.F., T. Karafet, A. Rasanayagam, E.T. Wood, T.K. Altheide, T. Jenkins, R.C. Griffiths, A.R. Templeton & S.L. Zegura, 1998. Out of Africa and back again: nested cladistic analysis of human Y-chromosome variation. *Molecular Biology and Evolution* 15, 427-41.

Malaspina, P., B.M. Ciminelli, L. Viggiano, C. Jodice, F. Cruciani, P. Santolamazza, D. Sellitto, R. Scozzari, L. Terrenato, M. Rocchi & A. Novelletto, 1997. Characterization of a small family (CAIII) of microsatellite-containing sequences with X-Y homology. *Journal of Molecular Evolution* 44, 652-9.

Malaspina, P., F. Cruciani, B.M. Ciminelli, L. Terrenato, P. Santolamazza, A. Alonso, J. Banyko, R. Brdicka, O. Garcia, C. Glaudiano, G. Guanti, K.K. Kidd, J. Lavinha, M. Avila, P. Mandich, P. Moral, R. Qamar, S.Q. Mehdi, A. Ragusa, G. Stefanescu, M. Caraghin, C. Tyler-Smith, R. Scozzari & A. Novelletto, 1998. Network analyses of Y-chromosomal types in Europe, North Africa and West Asia reveal specific patterns of geographical distribution. *American Journal of Human Genetics* 63, 847-60.

Malaspina, P., F. Cruciani, P. Santolamazza, A. Torroni, A. Pangrazio, N. Akar, V. Bakalli, R. Brdicka, J. Jaruzelska, A. Kozlov, B. Malyarchuk, S.Q. Mehdi, E. Michalodimitrakis, L. Varesi, M.M. Memmi, G. Vona, R. Villems, J. Parik, V. Romano, M. Stefan, M. Stenico, L. Terrenato, A. Novelletto & R. Scozzari, 2000. Patterns of male-specific inter-population divergence in Europe, West-Asia and North-Africa. *Annals of Human Genetics* 64, 395^112.

Mathias, N., M. Bayes & C. Tyler-Smith, 1994. Highly informative compound haplotypes for the human Y chromosome. *Human Molecular Genetics* 3, 115-23.

Santos, F.R., S.D.J. Pena & C. Tyler-Smith, 1995. PCR haplotypes for the human Y chromosome based on alphoid satellite DNA variants and heteroduplex analysis. *Gene* 165, 191-8.

Santos, F.R., A. Pandya, C. Tyler-Smith, S.D.J. Pena, M. Schanfield, W.R. Leonard, L. Osipova, M.H. Crawford & R.J. Mitchell, 1999. The central Siberian origin for Native American Y chromosomes. *American Journal of Human Genetics* 64, 619-28.

Schneider, S., J.-M. Kueffer, D. Roessli & L. Excoffier, 1997. *Arlequin ver.1.1: a Software for Population Genetic Data Analysis.* Geneva: Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Scozzari, R., F. Cruciani, P. Santolamazza, P. Malaspina, A. Torroni, D. Sellitto, B. Arredi, G. Destro-Bisol, G. De Stefano, O. Rickards, C. Martinez-Labarga, D. Modiano, G. Biondi, P. Moral, A. Dickers, D.C. Wallace & A. Novelletto, 1999. Combined use of biallelic and microsatellite Y-chromosome polymorphisms to infer affinities among African populations. *American Journal of Human Genetics* 65,829-46.

Whitfield, L.S., J.E. Sulston & P.N. Goodfellow, 1995. Sequence variation of the human Y chromosome. *Nature* 378, 379-80.