Brazilian Journal of Genetics, 19, 4, 665-670 (1996)
(Revista Brasileira de Genética)

MINI-REVIEW

Reading the human Y chromosome: the emerging DNA markers and human genetic history

Fabrício R. Santos and Chris Tyler-Smith

INTRODUCTION

The potential of Y-chromosomal DNA polymorphisms for identifying paternal lineages and contributing to studies of human genetic history has been appreciated for more than a decade (Casanova *et al.*, 1985). However, a lack of suitable markers has hindered progress in the field. Convenient markers which can be scored by PCR are only now becoming available (Roewer *et al.*, 1992; Seielstad *et al.*, 1994; Hammer and Horai, 1995; Whitfield *et al.*, 1995; Santos *et al.*, 1995a; Underhill *et al.*, 1996). Future progress in marker development is expected to be rapid and non-specialist labs will soon be able to add Y chromosomal DNA analysis to autosomal and mtDNA studies in order to obtain a more complete understanding of their population samples. This review will therefore describe the emerging markers and some of the ways in which they can be used.

Special features of the Y chromosome

The Y chromosome is haploid and most of it does not recombine at meiosis, so it is passed as a whole from father to son. Consequently, Y haplotypes can readily be constructed by combining the allelic states of multiple markers, and these haplotypes are stable except when changed by a mutation, which creates a new DNA polymorphism. Polymorphisms therefore provide information about the relationships of Y chromosomes, which will ultimately lead back (coalesce) to a single ancestor. The frequencies of Y haplotypes in populations can be used to determine the relationships among different people. Although haplotype information is not complicated by recombination, it can be complicated by recurrent mutation, so it is necessary to have a way of recognizing this when it occurs.

The Y chromosome is present in the population at one quarter of the frequency of autosomes. Its effective population size is further reduced by the high degree of variability in the number of offspring of males. The small number of Y chromosomes probably accounts for the low degree of sequence variability (Hammer, 1995; Whitfield *et al.*, 1995) and the slow progress in finding polymorphic markers.

Types of polymorphism

For many purposes, the most important characteristic of a marker is the number of times that it has arisen since the most recent common ancestor of existing Y chromosomes. We will distinguish here between *unique* polymorphisms that have arisen only once, and *recurrent* polymorphisms where the same allele has arisen independently on more than one occasion. How can we tell whether a polymorphism is unique or recurrent? The first guide is the molecular nature of the mutation. The insertion of an Alu sequence at any particular position is a very rare event, so the YAP (Y Alu Polymorphism) insertion (Hammer, 1994) can safely be assumed to have occurred only once. Point mutations are also rare (~10^{-9} per base per year; Hammer, 1995; Whitfield *et al.*, 1995), so most will have a unique origin. The frequencies expected for duplications and deletions are more difficult to predict,

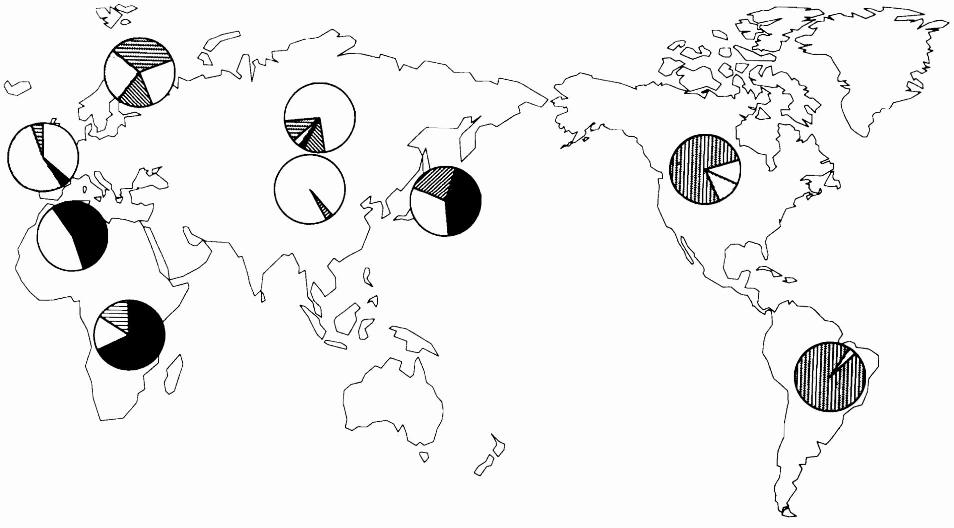
Department of Biochemistry, Oxford University, South Parks Road, Oxford OX1 3QU, UK. Fax: (44 1865) 275-283. E-mail: chris@bioch.ox.ac.uk. Send correspondence to C.T.-S.

but changes in the number of repeats in a microsatellite or minisatellite are expected to be frequent: an average rate of 2 x 10^{-3} per locus per generation for tetranucleotide microsatellites (Weber and Wong, 1993) and up to 7 x 10^{-2} per locus per generation for minisatellites (Buard and Vergnaud, 1994). With this information and the assumption that Y chromosomes do not recombine, Y haplotypes can be examined for evidence of recurrent mutations. As expected, haplotypes constructed using YAP and point mutations usually have simple relationships to one another (Seielstad *et al.*, 1994; Hammer and Horai, 1995), while haplotypes constructed using microsatellites show very complex relationships (Roewer *et al.*, 1996). These general features can then allow exceptions to be identified: recurrent point mutations, or microsatellite alleles that have arisen only once, for example.

Unique polymorphisms

The first point mutation to be described was 47z (*DXYS5Y*) (Nakahori *et al.*, 1989) and it illustrates the features typical of unique polymorphisms (Figure 1): the derived (i.e. non-ancestral) allele shows a simple relationship to the other markers, and has a restricted geographical distribution, being found at high frequency in Japan, at lower frequencies in Korea and Taiwan (Lin *et al.*, 1994), and not at all in the other populations tested. Similarly, YAP+ chromosomes (containing the Alu insertion) are found at high frequency in Africa and at low but detectable frequencies in much of Europe and Asia (Hammer, 1994). YAP illustrates two additional features of unique polymorphisms: firstly, its frequency is found to be unexpectedly high in another limited region (Japan), and secondly, a subset of YAP+ chromosomes (but no YAP- chromosomes) carry the derived sY81 G allele; these chromosomes have a more restricted geographical localization than YAP itself, and are largely confined to sub-Saharan Africa (Seielstad *et al.*, 1994). Likewise, Tat C allele chromosomes are found only in northern Asia and northern Europe (Zerjal *et al.*, unpublished results). 92R7 T allele chromosomes are quite widely distributed in Europe, parts of Asia, and North and South America (Mathias *et al.*, 1994 and unpublished data), while a subset of them, carrying the *DYS199* T allele, is confined to the Americas (Underhill *et al.*, 1996). In contrast, *SRY*-1532 A allele chromosomes (Whitfield *et al.*, 1995) are found both in Africa in association with the 92R7 C allele (group 6), and in Europe and Asia in association with the 92R7 T allele (group 3). Since the *SRY*-1532 G allele is also found in association with both 92R7 alleles, one of these point mutations must have occurred twice.

The geographical distributions shown in Figure 1 can be explained in simple ways: each mutation is likely to have occurred in the location where its derived allele is most frequent, and to have spread



Y Groups	1	3	18	2	5	16	4	8	6
	░	▤	▥	○	▨	▧	●	●	▤
loci									
92R7	1	1	1	0	0	0	0	0	0
SRY-1532	0	1	0	0	0	0	0	0	1
DYS199	0	0	1	0	0	0	0	0	0
47z	0	0	0	0	1	0	0	0	0
Tat	0	0	0	0	0	1	0	0	0
YAP	0	0	0	0	0	0	1	1	0
sY81	0	0	0	0	0	0	0	1	0

Figure 1 - Worldwide distribution of Y groups defined by seven unique or rare polymorphisms that can be detected using PCR. The allele state for each locus is: 92R7 (0 = C, 1 = T); *SRY*-1532 (0 = G, 1 = A); *DYS199* (0 = C, 1 = T); 47z (0 = *Stu*I site absent, 1 = *Stu*I site present); Tat (0 = T, 1 = C); YAP (0 = Alu sequence absent, 1 = Alu sequence present); sY81 (0 = A, 1 = G). The Y group frequencies are based on published work (Lin *et al.*, 1994; Mathias *et al.*, 1994; Hammer and Horai, 1995; Underhill *et al.*, 1996), two forthcoming papers (Zerjal *et al.*; Pandya *et al.*, and our unpublished results).

locally. Thus the 47z mutation probably arose in Southeast Asia, while the YAP insertion probably took place in sub-Saharan Africa. After YAP+ chromosomes had risen in frequency, and perhaps after some YAP+ chromosomes had left Africa, the sY81 A → G transition took place on a YAP+ background. The high frequency of YAP+ chromosomes in Japan can be explained in several ways: it could, for example, reflect an origin of the Japanese population from Asian ancestors carrying a high frequency of the marker who have subsequently been replaced on the mainland, or it could be due to genetic drift in a small founder population (Hammer and Horai, 1995). Thus, by correlating geographical distribution with Y haplotype relationships, it is possible to reconstruct aspects of population history. The number of markers available is still too small for a comprehensive reconstruction of world population movements, and those shown in Figure 1 are largely uninformative for some populations, such as the Chinese. More unique markers are needed, so that all chromosomes carry at least one derived marker. Markers which arose early and are widespread will be particularly useful, but will be fewer in number than those which have arisen recently.

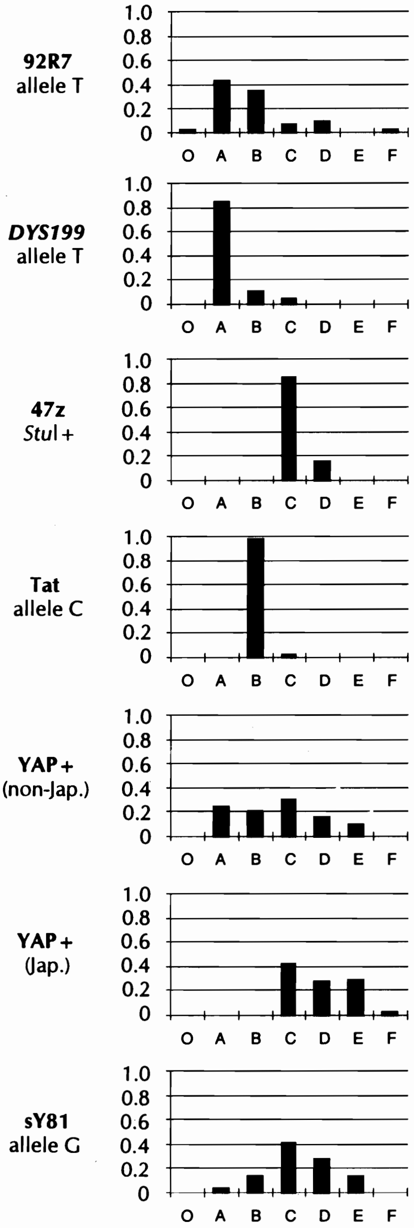


Figure 2 - *DYS19* allele distributions associated with unique polymorphisms. Alleles named according to Santos *et al.* (1996a): O = null allele, A = 186 bp, B = 190 bp, C = 194 bp, D = 198 bp, E = 202 bp and F = 206 bp. *DYS19* distributions in chromosomes containing YAP are divided into Japanese (Jap.) and non-Japanese (non-Jap.) populations. Some data for *DYS19* distributions are published elsewhere: *DYS199* (Underhill *et al.*, 1996), Tat (Zerjal *et al.*, unpublished results); 47z and YAP in Japanese (Hammer and Horai, 1995), YAP non-Japanese (Santos *et al.*, 1996b) and sY81 (Pandya *et al.*, unpublished results).

Recurrent polymorphisms

The most widely used recurrent polymorphisms are microsatellites. There have been extensive studies of the distribution of *DYS19* alleles in different populations (Santos *et al.*, 1996a) and the discriminating power of a combination of four loci is illustrated by the generation of 77 different haplotypes in a sample of 159 Dutch and Germans (Roewer *et al.*, 1996). We do not have space to review much of this work here, and will concentrate on one aspect: the relationship of unique polymorphisms to *DYS19* variability. *DYS19* is a tetranucleotide repeat with at least nine alleles (Roewer *et al.*, 1992; Santos *et al.*, 1993, 1996a), but sets of chromosomes defined by the unique polymorphisms display subsets of the *DYS19* alleles (Figure 2). In some cases (92R7, YAP, sY81) many alleles are found, but in other cases (*DYS199*, 47z, Tat) the number is greatly restricted. Each unique polymorphism arose on a single chromosome with one *DYS19* allele size, and the additional alleles have arisen by mutation. If all variants have an equal probability of surviving and the number of chromosomes expands, *DYS19* variability will increase with time and can potentially be used for dating (see below). Another recurrent polymorphism is the minisatellite MSY1, which consists of 60 to 100 copies of an A+T-rich 25 bp unit that itself exists in at least four forms (Jobling *et al.*, 1994). Population studies are eagerly awaited. A marker that combines some of the features of unique and recurrent mutations is the alphoid heteroduplex system (Santos *et al.*, 1995a). The amplification of 281-285 bp regions of variant alphoid satellite units generates patterns consisting of up to five heteroduplexes; 32 such patterns have now been found (unpublished observations). Each pair of heteroduplex bands probably has a unique origin, but loss of bands can be a recurrent event and simple patterns containing few bands or no heteroduplex can arise more than once by rare deletion events (Santos *et al.*, 1995b, 1996b). Nevertheless, the comparison of this system with other unique polymorphisms displays a remarkable association (Santos *et al.*, 1996b and unpublished results) and it is very useful for an initial survey of an unknown population when the amount of DNA is limited.

Population studies

Unique polymorphisms are useful for population studies when groups of people share the derived state of a marker: for example, the Japanese and Chinese populations both contain the 47z *Stu*I+ allele (Figure 1). Marker frequencies in several populations can be used to calculate population relationships, and it should now be possible to do this using multiple markers. Unique markers are not very useful when the populations contain only the ancestral allele: the African and South American populations, for example, would not be considered similar on the basis that they both contain 100% of the same (ancestral) 47z allele. Thus, different markers will be required in different parts of the world. However, it is still desirable to screen for all markers in all populations, so that unexpected migrations and recurrent mutations are detected. Microsatellites can also be used to compare populations. Populations with different microsatellite distributions are distinct (eg. Figure 3, A and B). However, populations with similar microsatellite distributions are not necessarily similar (eg. Figure 3, B and C). This is because convergent evolution can lead to different Y chromosomes (defined by their unique mutations) carrying the same distribution of microsatellite alleles. These difficulties are illustrated by two recent studies of the Finns. In one (Sajantila *et al.*, 1996) the analysis of *DYS19* and two non-informative markers led to the

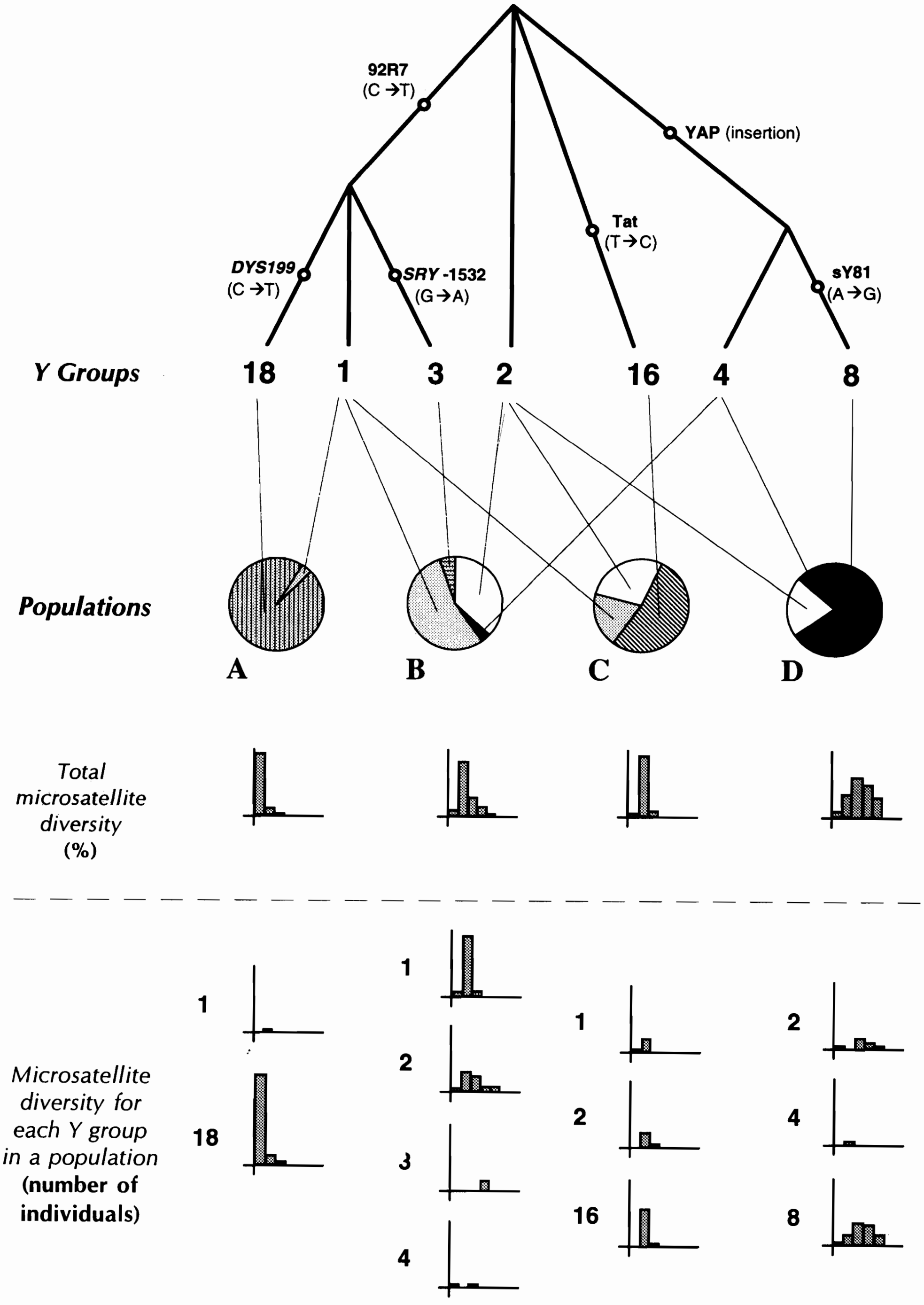

92R7
(C →T)
YAP (insertion)
Tat
(T→C)
DYS199
(C →T)
SRY -1532
(G→A)
sY81
(A→G)
Y Groups
18
1
3
2
16
4
8
Populations
A
B
C
D
Total microsatellite diversity (%)
Microsatellite diversity for each Y group in a population (number of individuals)
1
18
1
2
3
4
1
2
16
2
4
8

Figure 3 - Microsatellite (*DYS19*) variability and Y groups in selected populations (A = Amerindians, B = Caucasians from UK, C = Finns and D = Sub-Saharan Africans). The *DYS19* data (only considering alleles A to E) represent a compilation of published (Underhill *et al.*, 1996; Santos *et al.*, 1996b) and unpublished data (Zerjal *et al.*; Pandya *et al.*, unpublished results). The absence of recurrent mutation in the polymorphisms used for the classification in Y groups is assumed since some individuals were not typed with all systems.

conclusion that the Finnish population contained a subset of European Y chromosomes and that there had been a strong Y bottleneck. In the other, the analysis of the Tat point mutation (Zerjal *et al.*, unpublished results, and Figure 3) showed that the Finnish Y chromosomes differ from the other European populations tested and contain at least two frequent and distinct patrilineages that coincidentally are both associated to the same *DYS19* B allele. Thus, conclusions based on a single microsatellite locus can be misleading and additional informative markers should always be used.

Dating

One of the most useful pieces of information from any study of human evolution, whether paleontological or genetic, is a date. Our entire perspective can change when the date of a key event is revised. The possibility of using the Y chromosome to provide dates relevant to human history has therefore attracted attention. Two approaches have been used. In the first, the sequence diversity of human Y chromosomes has been compared with the sequence divergence between humans and chimpanzees in order to estimate a coalescence time for all human Y chromosomes. Times of 51,000-411,000 years (Hammer, 1995) and 37,000-49,000 years (Whitfield *et al.*, 1995) have been obtained when assuming a human-chimpanzee split at 5 million years ago. The wide range of dates reveals the uncertainties in these estimates, but despite this, provides support for a recent origin for modern humans.

The second approach to dating has been to measure the *DYS19* microsatellite diversity associated with a unique polymorphism such as the *DYS199* T allele, and use an average figure for the microsatellite mutation rate determined in modern families to estimate when the C → T transition took place (Underhill *et al.*, 1996). Unfortunately, when the published figure (Weber and Wong, 1993) was used this yielded a date of 2147 years ago, and it seems unlikely that the polymorphism could have spread through so much of the Americas in such a short time. The problem may be that mutation rates vary between loci and also depend on the allele size: it is notable that the *DYS199* T allele chromosomes carry a short *DYS19* array and might be expected to have a low mutation rate. Nevertheless, the approach is so attractive that it would be worthwhile to carry out the large amount of work necessary to measure size-specific, locus-specific mutation rates for the Y microsatellites, perhaps using large families or sperm DNA. Then, each locus can be used to estimate an independent date for each point mutation, and, if sufficient information about population structure is available, a detailed and calibrated reconstruction of the spread of human Y chromosomes throughout the world may be obtained.

The analysis of Y polymorphisms have been useful in the study of peopling of Japan (Hammer and Horai, 1995) and Amerindian migrations (Pena *et al.*, 1995; Santos *et al.*, 1995b, 1996c; Underhill *et al.*, 1996) but some major questions still remain unanswered. Do Y studies support an African origin for modern humans? Do African Y chromosomes show greater diversity than others? The markers now becoming available will help us to answer these questions.

ACKNOWLEDGMENTS

We thank the individuals who have donated DNA for their essential contribution to this work, Matt Hurles for his analysis of the 92R7 polymorphism, and Mark Jobling, Arpita Pandya and Sergio D.J. Pena for comments on the manuscript. Our research was supported by The Leverhulme Trust and the Cancer Research Campaign.

REFERENCES

Buard, J. and **Vergnaud, G.** (1994). Complex recombination events at the hypermutable minisatellite CEB1 (D2S90). *EMBO J. 13*: 3203-3210.

Casanova, M., Leroy, P., Boucekkine, C., Weissenbach, J., Bishop, C., Fellous, M., Purrello, M., Fiori, G. and **Siniscalco, M.** (1985). A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. *Science 230*: 1403-1406.

Hammer, M.F. (1994). 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. (1995). A recent common ancestry for human Y chromosomes. *Nature 378*: 376-378.

Hammer, M.F. and **Horai, S.** (1995). Y chromosomal DNA variation and the peopling of Japan. *Am. J. Hum. Genet. 56*: 951-962.

Jobling, M.A., Fretwell, N., Dover, G.A. and **Jeffreys, A.J.** (1994). Digital coding of human Y chromosomes: MVR-PCR at Y-specific minisatellites. *Cytogenet. Cell Genet. 67*: 390.

Lin, S.J., Tanaka, K., Leonard, W., Gerelsaikhan, T., Dashnyam, B., Nyamkhishig, S., Hida, A., Nakahori, Y., Omoto, K., Crawford, M.H. and **Nakagome, Y.** (1994). A Y-associated allele is shared among a few ethnic groups of Asia. *Jpn. J. Hum. Genet. 39*: 299-304.

Mathias, N., Bayés, M. and **Tyler-Smith, C.** (1994). Highly informative compound haplotypes for the human Y chromosome. *Hum. Mol. Genet. 3*: 115-124.

Nakahori, Y., Tamura, T., Yamada, M. and **Nakagome, Y.** (1989). Two 47z [*DXYS5*] RFLPs on the X and Y chromosomes. *Nucleic Acids Res. 17*: 2152.

Pena, S.D.J., Santos, F.R., Bianchi, N.O., Bravi, C.M., Carnese, F.R., Rothhammer, F., Gerelsaiknan, T., Munkhtuja, B. and **Oyunsuren, T.** (1995). A major founder Y-chromosome haplotype in Amerindians. *Nat. Genet. 11*: 15-16.

Roewer, L., Arnemann, J., Spurr, N.K., Grzeschik, K.-H. and **Epplen, J.T.** (1992). Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts. *Hum. Genet. 89*: 389-394.

Roewer, L., Kayser, M., Dieltjes, P., Nagy, M., Bakker, E., Krawczak, M. and **de Knijff, P.** (1996). Analysis of molecular variance (AMOVA) of Y-chromosome-specific microsatellites in two closely related human populations. *Hum. Mol. Genet. 5*: 1029-1033.

Sajantila, A., Salem, A.-H., Savolainen, P., Bauer, K., Gierig, C. and **Pääbo, S.** (1996). Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. *Proc. Natl. Acad. Sci. USA 93*: 12035-12039.

Santos, F.R., Pena, S.D.J. and **Epplen, J.T.** (1993). Genetic and population study of a Y-linked tetranucleotide repeat DNA polymorphism with a simple non-isotopic technique. *Hum. Genet. 90:* 655-656.

Santos, F.R., Pena, S.D.J. and **Tyler-Smith, C.** (1995a). PCR haplotypes for the human Y chromosome based on alphoid satellite variants and heteroduplex analysis. *Gene 165*: 191-198.

Santos, F.R., Hutz, M.H., Coimbra, C.E.A., Santos, R.V., Salzano, F.M. and **Pena, S.D.J.** (1995b). Further evidence for the existence of major founder Y chromosome haplotype in Amerindians. *Braz. J. Genet. 18*: 669-672.

Santos, F.R., Gerelsaikan, T., Munkhtuja, B., Oyunsuren, T., Epplen, J.T. and **Pena, S.D.J.** (1996a). Geographic differences in the allele frequencies of the human Y-linked tetranucleotide polymorphism *DYS19*. *Hum. Genet. 97*: 309-313.

Santos, F.R., Bianchi, N.O. and **Pena, S.D.J.** (1996b). Worldwide distributions of Y chromosome haplotypes. *Genome Res. 6*: 601-611.

Santos, F.R., Rodriguez-Delfin, L., Pena, S.D.J., Moore, J. and **Weiss, K.M.** (1996c). North and South Amerindians may have the same major founder Y chromosome haplotype. *Am. J. Hum. Genet. 58*: 1369-1370.

Seielstad, M.T., Herbert, J.M., Lin, A.A., Underhill, P.A., Ibrahim, M., Vollrath, D. and **Cavalli-Sforza, L.L.** (1994). Construction of human Y-chromosomal haplotypes using a new polymorphic A to G transition. *Hum. Mol. Genet 3*: 2159-2161.

Underhill, P.A., Jin, L., Zemans, R., Oefner, P.J. and **Cavalli-Sforza, L.L.** (1996). A pre-Colombian Y chromosome-specific transition and its implications for human evolutionary history. *Proc. Natl. Acad. Sci. USA 93*: 196-200.

Weber, J.L. and **Wong, C.** (1993). Mutation of human short tandem repeats. *Hum. Mol. Genet. 2*: 1123-1128.

Whitfield, L.S., Sulston, J.E. and **Goodfellow, P.N.** (1995). Sequence variation of the human Y chromosome. *Nature 378*: 379-380.