Submitted, *Current Anthropology*

Human genetic and linguistic diversity: Continental patterns and time depth

Daniel Nettle
Departments of Biological Sciences and Psychology
Open University
Walton Hall
Milton Keynes, UK
MK7 6AA

E-mail: D.Nettle@open.ac.uk

Key words: biological anthropology, linguistic diversity, genetic diversity, anatomically modern humans

Theory predicts that, other things remaining constant, genetic diversity within a population will increase with time, until an equilibrium dictated by population size, selection and the mutation rate has been reached (Page and Holmes 1998: 109; Seielstad et al. 1999: 563). Populations do not leap into being from a homogenous starting point, often inheriting diversity from their precursors, but major events such as speciation or the founding of a new continent by a small group can act as bottlenecks, drastically reducing diversity in the short term. Whilst factors other than time, such as migration, population size, and selection, can affect the subsequent accumulation of diversity, it is a reasonable hypothesis that internal diversity will roughly correlate with time depth. This is the interpretation most often given to the relatively high human genetic diversity in Africa compared to the other continents (Cann, Stoneking and Wilson 1987; Cavalli-Sforza, Piazza and Menozzi 1994; Vigilant et al. 1991; Bowcock et al. 1994; Seielstad et al. 1999; Ingman et al. 2000; Jorde et al. 2000; Thomson et al. 2000), though other interpretations are possible (Relethford 1995).

It has been proposed that linguistic and genetic diversity have evolved in tandem in the diaspora of modern humans (Cavalli-Sforza et al. 1988). This proposal has caught the scientific imagination sufficiently strongly for there to be talk of a 'new synthesis' of genetics, archaeology and linguistics, to give a unified account of human history (Renfrew 1992). If the hypothesis of tandem evolution is true, then one might predict a triple correlation between time depth, genetic diversity and linguistic diversity. In this paper, I investigate this hypothesis by correlating measures of genetic and linguistic diversity with archaeological estimates of the time depth of the population for all the major continents.

Methods

Genetic Diversity

Published continental data from five genetic systems were assembled. The choice of which data sets to use was determined simply by the availability within each study of the some data for all of the five major continents. The types and sources of data are shown in table 1. Sample sizes and structures vary for the several sources, as do the

types of data. The mtDNA data are mean pairwise differences based on complete sequencing of the mtDNA molecule (thus inter-individual differences). All other data are diversity measures derived from gene or haplotype frequencies (and thus, between sub-population diversities).

Linguistic Diversity

Linguistic diversity can be characterised at one level by simply counting the number of languages in a given area. This level of diversity reflects social, political and ecological factors operating over very recent time scales (Nettle 1999). To make inferences about the more remote past, diversity needs to be studied at a deeper level, that of the number of families into which contemporary languages are grouped. The global distribution of linguistic diversity at this deeper level was worked out by Nichols (1990a,1992). Nichols' basic unit is the stock, the maximally deep language family reconstructable by the comparative method of historical linguistics. Stocks are assumed to all represent entities of similar evolutionary depth. To compensate for the different sizes, ecologies and population densities of the continents, a measure of relative linguistic diversity has been calculated. This is the number of stocks per thousand languages spoken (S/L). Many languages all belonging to a few stocks gives a low value, whilst the same number of languages belonging to many distinct stocks gives a higher one.

Time Depth

Time depths of the populations are arrived at in the following ways. In Africa, the earliest anatomically modern human remains date from 130,000 bp (Day and Stringer 1991) . Mitochondrial coalescent analysis of the whole human population provides central date estimates between this time and 200,00 bp (Cann, Stoneking and Wilson 1987; Vigilant et al. 1991). Though the African population may retain some early diversity, it was very small until 100,000 bp or so (Harpending 1994). Thus, we can take 130,000 years as a time depth yardstick for African (and background human) diversity. This date is not critical; the important consideration for present purposes is

that the Africa population is much older than any other on earth in terms of time since a known dramatic bottleneck.

Australia and New Guinea were settled around 60,000 Kyr (Roberts et al, 1994; Thorne et al. 1999), through a migration East through South Asia. A separate dispersal North from Africa into the Middle East begins at around 45 Kyr, with spread into Europe shortly afterward (Mellars 1993). Thus the time depth vales for Europe (40 Kyr) and Oceania (60 Kyr) are obvious. Asia is more problematic. Though the 60Kyr eastward expansion went through South Asia, contemporary Asian populations are morphologically closer to the Levantine/Caucasian group than to Australian aborigines (Lahr 1996). The exceptions are outlier populations in India and island Southeast Asia. Thus much of the genetic diversity in Asia may come from the 45Kyr Middle Eastern expansion. Nonetheless, some genetic contribution from the earlier Asians is likely, not just through the persistence of outlier populations but also through intermarriage. Therefore, in terms of the origin of diversity, Asia must be assigned the first, 60 Kyr time depth.

Controversy remains over the exact date of the settlement of the Americas. The traditional archaeological ('Clovis') date of 11 Kyr must be revised earlier in the light of recent dated finds (Dillehay 1996), but this revision may not be very substantial, and climatic factors make a date as far back as the glacial maximum at 18 Kyr unlikely. I have thus taken 15 Kyr as the value here. The precise number is less important than the fact that the Americas is a much younger population than the African or Asian ones. Several distinct mtDNA lineages were involved in the founding (Ward 1999), and there has been some subsequent migration from Asia (such as the Na-Dene and Eskimo groups). Nonetheless, it can be assumed in view of the geographic difficulty of the migration route, through Eastern Siberia, Beringia, and the North American ice sheets, that the settlement of the Americas constituted a significant population bottleneck.

Results

The data are presented in table 2. Genetic diversity is highest in the oldest continent Africa for mtDNA, Y chromosome, and microsatellites, but not for classical markers

or RFLPs. Minimum diversity is in the Americas for Y chromosome and microsatellites. The relationships between genetic diversity and time depth are shown in figure 1. The plots are suggestive of increasing diversity with increasing time depth for mtDNA, Y chromosome and microsatellites, but not for classical markers or RFLPs. None of the rank correlations between time depth and genetic diversity is individually significant, but it is extremely difficult to obtain significant rank correlations with only five data points. However, when the three correlation coefficients from mtDNA, Y chromosome and microsatellites are considered collectively using Fisher's procedure for combining probabilities (Sokal and Rohlf 1981), they differ significantly from 0 ($Chi^2 = 13.71$, d.f. = 6, $p < 0.05$).

The plot of linguistic diversity against time depth, by contrast, suggests a negative relationship (figure 2; $r_s = -0.821$, $p = 0.089$). The relationship between genetic and linguistic diversity is thus negative (for Y chromosome, $r_s = -1$, $p < 0.001$; for microsatellites, $r_s = -0.9$, $p < 0.05$; for the other systems, $r_s = -0.5$, n.s.). All five correlations between genetic and linguistic diversity are negative, and, considered together, they differ significantly from 0 ($Chi^2 = 26.04$, d.f. = 10, $p < 0.01$).

Discussion

Few though the data are, the results suggest that genetic diversity does increase with the time depth of the population Positive relationships between time depth and genetic diversity hold for mtDNA, Y chromosome and microsatellites, but not for nuclear RFLPs, or classical markers. Part of the reason for this may be that the latter two systems were ascertained as polymorphic in the European population. Thus, there will be a bias inflating the apparent diversity of Europeans and populations close to them (Bowcock et al. 1994; Jorde et al. 2000). The mtDNA data is based on the sequence of the entire molecule, and so does not suffer such problems. Both the Y chromosome and the microsatellite data show high absolute levels of diversity, which is sufficient to overcome any possible ascertainment bias (Rogers and Jorde 1996). Ascertainment bias may not be the only factor, though, since direct comparison of some non-coding autosomal sequences does not reveal greatly increased nucleotide diversity in Africa (Zhao et al. 2000). However, these sequences have low mutation

rates compared to microsatellites or mtDNA, and, since they recombine, have effective population sizes four times that of non-recombining systems. The former factor means they will be slow to register evolutionary divergence, whereas the latter means that they will be less affected by bottlenecks than mtDNA and the Y chromosome.

Linguistic diversity shows a completely different pattern, being at its maximum in the youngest continent the Americas and low in Africa. The decrease with time appears paradoxical, since it must be the case that diversity is very low when a continent is first settled. This has been addressed elsewhere (Nettle 1999; Nichols 2000). After an initial diversification phase, the principal evolutionary force in linguistic diversity becomes the extinction of lineages, as small groups become assimilated into larger ones in the context of population growth and trends towards larger-scale forms of political and economic organisation. The initial diversification phase could be very fast. When a continent is first settled, there are a large number of empty geographical niches for small communities to occupy. The first radiation will thus produce a star-shaped phylogeny and as many branches as there are geographical niches. The opportunity for new lineage fissions after this time is limited, and since linguistic extinction is thereafter the dominant force, diversity can only decline. This explains the observed negative relationship with time depth.

These results have obvious implications for the hypothesis of Cavalli-Sforza et al. (1988) that linguistic and genetic diversity evolve in tandem. Their often-quoted comparison of family trees appears to show congruence between distributions of linguistic and genetic diversity, whereas the present results show quite the opposite. For several reasons, it seems likely that the apparent congruence of the distributions in Cavalli-Sforza et al. (1988) is artifactual, as has been argued elsewhere (Bateman et al. 1990; Nichols 1990b; McMahon and McMahon 1995; Sims-Williams 1998). First, precise reconstruction of the linguistic tree has not proved possible back to a global root, as linguistic information erodes too fast. The tree used by Cavalli-Sforza et al. (1988), following Ruhlen (1987), is in fact a geographical ordering of nodes which are mainly loose geographical groups of languages rather than demonstrated phylogenetic entities (Bateman et al. 1990; Nichols 1990b; Dixon 1997; Campbell 1998). These groupings mask the real distribution of diversity. South America , for example, is

shown as a single sub-node of ‘Amerind’, whereas in fact no accepted phylogenetic reduction of these languages has been published, and systematic linguistic comparison suggests that South America contains at least 90 major families, far more than Africa and Eurasia combined (Nichols 1990a,1992). The nodes ‘Indo-European’ and ‘Amerind’are shown as entities of the same depth in the tree, but they are in reality not remotely comparable. Indo-European is a readily reconstructable language family. Amerind is an as yet unsubstantiated macro-grouping of well over a hundred families which if related at all are so remote that the evidence is very cryptic, and many of which are *internally* more divergent than Indo-European.

The Cavalli-Sforza et al. ‘tree’ contains only limited heirarchical structure, with many of the branchings being multiple, including an initial 11 way split. Cavalli-Sforza et al. exploit the mobile properties of trees and choose to display the ordering of branches from the initial 11-way split which is most congruent with geography (and thus the genetic tree), which unjustifiably inflates the apparent similarity. The primary determinant of the structure of genetic tree is simply geographic distance, because of isolation by distance effects (Cavalli-Sforza, Piazza and Menozzi 1994). Thus both ‘trees’ are really geographical arrays, and their correlation fairly trivial.

This does not mean that there are no levels at which genetic and linguistic diversity coevolve. Linguistic and genetic boundaries often coincide (Barbujani and Sokal 1990; Barbujani 1991; Dupanloup et al. 2000), since, locally, similar geographical and cultural processes impede gene flow and linguistic interaction. However, this finding does not ‘scale up’ to the global distribution of diversity. The dynamics of transmission - the rate of change, the rate of extinction, the significance of partial cultural isolation - are simply not the same in the two cases. This does not mean that the ‘new synthesis’ is doomed. Linguistic diversity, like genetic diversity, furnishes a set of evolutionary markers that reflect population processes (Nettle 1999). There is no reason why the information from these markers cannot ultimately be integrated with information from genetic markers. However, not all markers are informative at the same level or are subject to the same evolutionary dynamics. Geneticists increasingly understand how every genetic system has a unique trajectory, the outcome of its particular combination of mutation rate, selection, and mode of

transmission. Likewise, the new synthesis will require better models of the spatial and temporal trajectories of the very different evolutionary systems it covers.

References

BARBUJANI G. 1991 What do languages tell us about human microevolution?, *Trends in Ecology and Evolution* 6: 151-6

BARBUJANI, G., AND R.R. SOKAL. 1990. Zones of sharp genetic change in Europe are also linguistic boundaries. *Proceedings of the National Academy of Sciences of the U.S.A.* 87: 1816-1819.

BATEMAN R., I. GODDARD, R. O'GRADY, V.A. FUNK, R. MOOI, W.J. KRESS AND P. CANNELL. 1990. Speaking with forked tongues: the feasibility of reconciling human phylogeny and the history of language. *Current Anthropology* 31: 1 – 24.

BOWCOCK, A.M., A. RUIZ-LINARES, A. TOMFOHRDE, E. MINCH, J.R. KIDD, AND L.L. CAVALLI-SFORZA. 1994. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 368: 455-457.

CAMPBELL, L. 1998. *Historical Linguistics*. Edinburgh: Edinburgh University Press.

CANN R.L., M. STONEKING AND M. WILSON. 1987. Mitochondrial DNA and human evolution. *Nature* 325: 31-6.

CAVALLI-SFORZA L.L., PIAZZA A., MENOZZI P. AND MOUNTAIN J. 1988. Reconstruction of human evolution : bringing together genetic, archaeological and linguistic data. *Proceedings of the National Academy of Sciences of the USA* 85: 6002 - 6.

CAVALLI-SFORZA L.L., A. PIAZZA AND P. MENOZZI. 1994. *The History and Geography of Human Genes*. Princeton: Princeton University Press.

DAY, M.H. AND C.B. STRINGER. 1991. Les restes craniens d'Omo-Kibish et leur classification a l'interieur du genre *Homo*. *L'anthropologie* 95:573-594.

DILLEHAY, T.D. 1996. *Monte Verde: A Late Pleistocene Settlement in Southern Chile. Volume 2: The Archaeological Context*. Washington, DC: Smithsonian Institute.

DIXON, R.M.W. 1997. *The Rise and Fall of Languages*. Cambridge: Cambridge University Press.

DUPANLOUP DE CEUNINCK, I. ET AL. 2000. Inferring the impact of linguistic boundaries on population differentiation: Application to the Afro-Asiatic-Indo-European case. *European Journal of Human Genetics* 8: 750-6.

GRIMES, B.F. 2000. *Ethnologue: The World 's Languages. 14th Edition.* Dallas: Summer Institute of Linguistics.

HAMMER, M.F. et al. (1996). The geographic distribution of human Y chromosome variation. *Genetics* 145: 787-805.

HARPENDING, H. 1994. Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. *Human Biology* 66: 591-600.

INGMAN, M. et al. 2000. Mitochondrial genome variation and the origin of modern humans. Nature 408: 708-13.

JORDE, L. B., W. S. WATKINS, M. J. BAMSHAD, M. E. DIXON, C. E. RICKER, M. T. SEIELSTAD, AND M. A. BATZER. 2000. The distribution of human genetic diversity: A comparison of mitochondrial, autosomal, and Y-chromosome data. *American Journal of Human Genetics* 66: 979-988.

LAHR, M.M. 1996. *The Evolution of Human Diversity*. Cambridge: Cambridge University Press.

MCMAHON, A.M.S. AND R. MCMAHON. 1995. Linguistics, genetics and archaeology: Internal and external evidence in the Amerind controversy. *Transactions of the Philological Society* 93: 125-226.

MELLARS, P.A. 1993. "Archaeology and the population-dispersal hypothesis of modern human origins in Europe, " in *The Origins of Modern Humans and the Impact of Chronometric Dating*. Edited by M. J. Aitken, C. B. Stringer, and P. A. Mellars, pp. 196-216. Princeton, Princeton University Press.

NETTLE, D. 1999. *Linguistic Diversity*. Oxford: Oxford University Press.

NICHOLS, J. 1990a. Linguistic diversity and the first settlement of the New World. *Language* 66: 475-521.

NICHOLS, J. 1990b. More on human phylogeny and linguistic history. *Current Anthropology* 31:313-4.

NICHOLS, J. 1992. *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press.

NICHOLS, J. 2000. "Estimating dates of early American colonization events" In *Time Depth in Historical Linguistics,* volume 2. Edited by C. Renfrew, A. McMahon and L. Trask, pp. 643-654. Cambridge: The McDonald Institute for Archaeological Research.

PAGE, R. AND E. HOLMES. 1998. *Molecular Evolution: A phylogenetic approach.* Oxford: Blackwell.

RELETHFORD, J.H. 1995. Genetics and modern human origins. *Evolutionary Anthropology* 4: 53-63.

RENFREW, C. 1992. Archaeology, genetics and linguistic diversity. *Man* 27, 445-78.

ROBERTS, R.G., R. JONES, N.A. SPOONER, M.J. HEAD, A.S. MURRAY, AND M.A. SMITH. 1994. The human colonisation of Australia: optical dates of 53,000 and 60,000 years bracket human arrival at Deaf Adder Gorge, Northern Territory. *Quaternary Science Reviews* 13,575-586.

ROGERS, A.R. AND L.B. JORDE. 1996. Ascertaiment bias in estimates of average heterozygosity. *Human Biology* 67:1-36.

RUHLEN, M. 1987. *A Guide to the World's Languages. Volume 1: Classification*. London: Edward Arnold.

SEIELSTAD, M. ET AL. 1999. A view of modern human origins from Y chromosome microsatellite variation. *Genome Research* 9: 558-567.

SIMS-WILLIAMS P. 1998. Genetics, linguistics and prehistory : thinking big and thinking straight. *Antiquity* 72: 505 – 27.

SOKAL, R.R. AND ROHLF, F.J. 1981. *Biometry: Principles and Practice of Statitiscs in Biological Research.* 2nd edition. Oxford: W.H. Freeman.

THOMSON, R., J. K. PRITCHARD, P. D. SHEN, P. J. OEFNER, AND M. W. FELDMAN. 2000. Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data. *Proceedings of the National Academy of Sciences of the United States of America* 97: 7360-7365.

THORNE, A, R. GRUN, G. MORTIMER, N.A. SPOONER, J.J. SIMPSON, M. NMCCULLOCH, L. TAYLOR, AND D. CURNOE. 1999. Australia's oldest human remains: age of the Lake Mungo 3 skeleton. *Journal of Human Evolution* 36: 591-612.

VIGILANT, L., M. STONEKING, H. HARPENDING, K. HAWKES, AND A. WILSON. 1991. African populations and the evolution of human mitochondrial DNA. *Science* 253: 1503-1507.

WARD, R. 1999. “Language and genes in the Americas,” in *The Human Inheritance: Genes, Language and Evolution*. Edited by B. Sykes, pp. 135-158. Oxford: Oxford University Press.

ZHAO, Z. et al. 2000. Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. *Proceedings of the National Academy of Sciences of the USA* 97: 11354-8.

System	Data type	Measure	Source
mtDNA	Pairwise differences, complete sequence	MPSD	Ingman et al. (2000)
Y chromosome	Frequencies of 27 combination haplotypes using 7 markers	Nei diversity measure *100	Hammer et al. (1997)
Microsatellites	Frequencies, 30 microsatellites	Mean heterozygosity *100	Bowcock et al. (1994)
Classical markers	Frequencies, 110 protein markers	Mean heterozygosity *100	Bowcock et al. (1994)
RFLPs	Frequencies	Mean heterozygosity *100	Bowcock et al. (1994)
Languages	Estimate by continent	Total	Grimes (2000)
Stocks	Estimate by continent	Total	Nichols (1992)
S/L	Stocks per thousand languages	-	-
Time	See text	Thousand years	

Table 1. Types and sources of data on genetic and linguistic diversity

Continent	mtDNA	Y Chromosome	Microsats	Classical	RFLPs	Languages	Stocks	S/L	Time
Africa	76.7†	88†	80.7†	16.3	29.7	2011	20	9.9§	130
Asia	36.7	78	68.5	18.9	32.7	2165†	22	10.2	60
Europe	24.7§	74	73	20.2†	37.9†	225§	6§	26.7	40
Oceania	41.8	72	63.6	13.7§	27.5§	1302	46	35.3	60
Americas	36.2	56§	58.8§	15.5	29.3	1000	157†	157†	15

† Maximum diversity
§ Minimum diversity

Table 2. Continental comparisons for genetic diversity, linguistic diversity and time depth

System	Time Depth	Linguistic	Classical	RFLPs	Microsat.	Y Chrom.
mtDNA	0.872	-0.500	-0.500	-0.500	0.300	0.500
	(0.054)	(0.391)	(0.391)	(0.391)	(0.624)	(0.391)
Y Chrom.	0.821	-1.000	0.500	0.500	0.900	
	(0.089)	(0.001)	(0.391)	(0.391)	(0.037)	
Microsat.	0.667	-0.900	0.600	0.600		
	(0.219)	(0.037)	(0.285)	(0.285)		
RFLPs	-0.051	-0.500	1.000			
	(0.935)	(0.391)	(0.001)			
Classical	-0.051	-0.500				
	(0.935)	(0.391)				
Linguistic	-0.821					
	(0.089)					

Table 3. Spearman rank correlation coefficient between the genetic and linguistic diversity data and time depth. p values are shown in brackets.

Figure Captions

Figure 1. Plots of diversity against time depth for the five genetic systems.

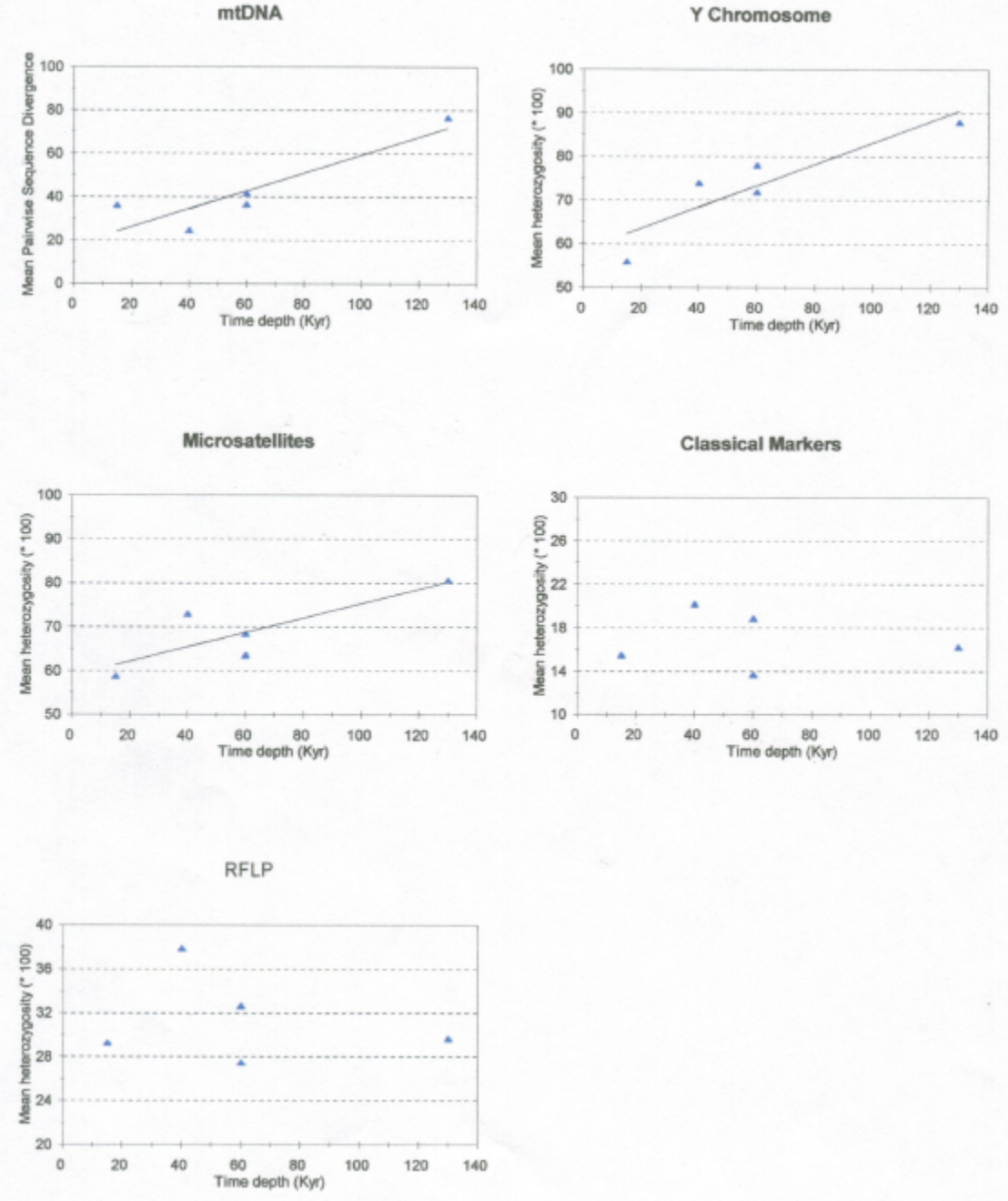

Figure 2. Plot of linguistic diversity against time depth.