Int J Leg Med (1993) 106:85–90

International Journal of
Legal Medicine

The application of mitochondrial DNA typing to the study of white Caucasian genetic identification

Romelle Piercy, K.M. Sullivan, Nicola Benson, and P. Gill

Central Research and Support Establishment, Forensic Science Service, Aldermaston, Reading, Berkshire, RG7 4PN, UK

Received April 29, 1993 / Received in revised form June 2, 1993

Summary. Mitochondrial DNA (mtDNA) from 100 unrelated British White Caucasians was extracted, amplified and directly sequenced. Sequences of approximately 800 nucleotides were obtained from 2 hypervariable segments within the non-coding region of the mitochondrial genome. A total of 91 different sequences were observed with an average nucleotide diversity of 1.1%. The most diverse pair of sequences differed at 3.6% of their nucleotide (nt) sites. Comparison to a consensus reference sequence showed that each region was polymorphic to a similar extent. Different methods of genetic analysis were used to examine the variation in each region, including pairwise comparisons, which demonstrated that although the data did not fit a Poisson distribution, the fit was closer to a Negative Binomial distribution.

Key words: Mitochondrial DNA – White Caucasian – Hypervariable sites – Pairwise distribution

Zusammenfassung. Die mitochondriale DNA (mtDNA) von 100 unverwandten britischen Kaukasiern wurde extrahiert, amplifiziert und direkt sequenziert. Die Sequenzen von ungefähr 800 Nucleotiden wurden von 2 hypervariablen Segmenten innerhalb der nicht-kodierenden Region des mitochondrialen Genoms erhalten. Insgesamt wurden 91 unterschiedliche Sequenzen beobachtet mit einer durchschnittlichen Nukleotid-Diversifität von 1,1%. Die meisten unterschiedlichen Sequenz-Paare differierten an 3,6% ihrer Nukleotid-(nt)-Positionen. Ein Vergleich mit einer Konsensus-Sequenz zeigte, daß jede Region bis zu einem ähnlichen Ausmaß polymorph war. Unterschiedliche Methoden der genetischen Analyse wurden angewandt, um die Variation in jeder Region zu untersuchen, unter Einschluß paarweiser Vergleiche. Diese Untersuchung zeigte, daß die Daten nicht einer Poisson-Verteilung entsprachen, sie paßten eher zu einer negativen Binomial-Verteilung.

Correspondence to: R. Piercy

Schlüsselwörter: Mitochondriale DNA – Weiße Kaukasier – Hypervariable Orte – Paarweise Verteilung

Introduction

For some years many different methods have been developed in order to characterize mtDNA [1–7]. Primarily this has been attempted to try to link diseases to specific mitochondrial defects [8–11] and also, because of the maternal inheritance exhibited, to trace the phylogenetic history of the human population and examine population structuring [5, 7, 12–31]. The genetic identification involved, however, may also have applications in other fields especially, for example, in forensic biology where positive identification of human remains and/or suspect traces are required to establish guilt or innocence of individuals thought to be responsible for a crime [3, 6]. However, in some circumstances identification has been difficult because of decomposition or degradation of soft tissues or presence of limited amounts of material. Often the only samples available may be bone, teeth or hair which contain reduced quantities of genomic DNA, but significant amounts of mtDNA.

Although restriction analysis has previously been utilised, PCR and direct sequencing of mtDNA enables a more rapid and reliable analysis of the polymorphic segments [2, 4]. Differences between individuals in these polymorphic sequences may, in principle, permit identification from samples containing only a few copies of the DNA of interest. The control region of mtDNA, which includes the origin of H strand replication, the displacement (D) loop and both origins of transcription, is the most variable region of the human mtDNA genome [32]. The majority of the polymorphisms were concentrated in 2 hypervariable segments, one of which lies around the origin of replication, the other lies within the D loop itself [12, 13, 33]. Both hypervariable segments have been analysed in this study in order to investigate genetic identification within the British White Caucasian population.

Materials and methods

The entire mitochondrial non-coding region was amplified using primers L15926 and H580 [1,2] in equal concentrations. Further amplifications of the 2 variable segments of this non-coding region were then carried out using an aliquot of the first amplification products as template and primer pairs L15997 with H16401, and L29 with H408 [1]. Amplification products were sequenced using solid phase separation of strands and Sequenase sequencing [2]. The regions were sequenced in both orientations in order to verify the accuracy of base-calling. In general, 403 nucleotides in region 1 and 380 nucleotides in region 2 were determined.

DNA extraction. Blood samples were obtained from 100 unrelated white Caucasians from 6 different regions in England and Wales. Total DNA was extracted from these samples as follows: in a 1.5 ml Eppendorf tube 200 µl liquid blood was added to an equal volume of extraction buffer (Tris base 1.21 g/l, EDTA 3.72 g/l, NaCl 5.84 g/l) containing 2% sodium dodecyl sulphate (SDS), dithiothreitol (DTT, 6 m*M* final concentration) and proteinase K (1.5 mg/ml). This mixture was incubated overnight at 37°C, extracted twice with phenol and precipitated using absolute ethanol at −20°C for 30 mins. The DNA was pelleted by spinning in a microcentrifuge for 10 mins, washed in 70% ethanol, vacuum-dried briefly and resuspended in 50 µl distilled water (dH_2O). The amount of DNA present was quantitated fluorimetrically with a Hoefer TKO 100 Mini fluorimeter.

PCR amplification of mtDNA sequences. PCR amplification was conducted in 2 stages: in the first PCR reaction the DNA template was amplified using a 25 µl reaction containing 10 ng of extracted DNA, primers (L15926 and H580 [1, 2] at 1 µ*M*, 200 µ*M* of each dNTP, 1.25U Taq polymerase (Perkin-Elmer Cetus) and 1 × amplification (PARR) buffer (Cambio) to produce a 1.3 kb PCR product encompassing the entire mtDNA control region. An aliquot of 0.5 µl of the DNA product was used without purification for subsequent nested PCR reactions in a 50 µl volume containing the following primers: biotinylated L15997 with M13(−21)-H16401 for L strand of variable region, 1, M13(−21)-L15997 with biotinylated H16401 for H strand of region 1, biotinylated L29 with M13(−21)-H408 for L strand of variable region 2 or M13(−21)-L29 with biotinylated H408 for H strand of region 2 at 0.5 µ*M* final concentration, 20 µ*M* final concentration each dNTP, 2.5U Taq polymerase, 1 × PARR buffer. This reaction generated biotinylated PCR products in which the M13(−21) universal sequencing primer sequence was incorporated at one end.

PCR was conducted in a thermal cycler (480 model, Perkin-Elmer Cetus) using 25 cycles, denaturation at 94°C for 45 secs, annealing at 50°C for 30 secs and elongation at 72°C for 5.5 mins for the first round amplification. Second round nested amplifications comprised 25 cycles of denaturation at 94°C for 45 secs, annealing at 50°C for 30 secs and elongation at 72°C for 3 mins. The quality of the PCR reaction was determined by electrophoresis of 5 µl of the final reaction mixture on gels containing 3% NuSieve GTG (FMC BioProducts) in TBE buffer (Tris base 15.75 g/l, Boric acid 4.64 g/l and EDTA 0.93 g/l).

Sequencing mtDNA. The biotinylated PCR product was attached to 30 µl streptavidin coated Dynabeads (Dynal A.S., Oslo, Norway) by incubating in LiCl buffer at 48°C for 15 mins. The non-biotinylated strand was removed by denaturing the complex in 0.15 *M* NaOH for 4 mins at room temperature. The resulting Dynabead/single-stranded PCR product complex was stabilised with the use of a magnet and the supernatant containing the eluted strand was removed. The Dynabead complex was resuspended in 16 µl dH_2O and used directly in sequencing reactions using the Sequenase Dye Primer Sequencing kit (Applied Biosystems) and employing the M13 universal sequencing primer. Electrophoresis and sequence analysis were performed with an Applied Biosystems Model 373A Automated DNA Sequencer. Comparisons of DNA sequences were carried out using a SeqEd package (Applied Biosystems).

Results and discussion

Comparison with the reference Anderson sequence

Alignments were made with the original Anderson sequence [32] and all differences were noted (Fig. 1). The Anderson sequence was atypical at 3 positions: nt 263, guanine was normally present instead of adenine; at nt 302–308 the majority of sequences contained 8 cytosine nucleotides instead of the 7 reported by Anderson; and at nt 310–314 6 cytosine nucleotides, not five, were usually present. Thus a modified reference sequence incorporating these changes was used in all further analyses. This consensus sequence may be representative of the ancestral state of white Caucasian mtDNA.

Comparisons showed that there were more sites of sequence polymorphism in the first region than in the second (76:48). However, the frequencies of polymorphism of individual sites were lower in the first region, with the result that each segment had equivalent levels of polymorphism (Figs. 1 and 2). This contrasts with previously published data [1, 12, 13, 33], where only a limited sample size (7–14) was used, in which the second region appeared to be less variable than the first.

However, in agreement with previously published data [12–17, 32, 33] the distribution showed a large bias towards transitional changes with a transition : transversion ratio of 42.5 : 1. This excess level of transitions may be indicative that mispairing during replication is the major source of spontaneous mutations within the mitochondria [18]. Although these previous publications suggested that the majority of observed transitions were pyrimidine, our data showed that the bias (80% observed) was only present in the first variable segment, with the second segment showing approximately equal levels of pyrimidine and purine transitions.

Genetic analysis of white Caucasian mtDNA data

Genetic identity testing using $p = \Sigma \times^2$ (where × was the frequency of mitochondrial genotype) for the two regions together, gave values of 0.026 (sequence specific oligonucleotide typing, using a database of 525 individuals from 5 ethnic groups) and 0.053 (sequencing, using samples obtained from 9 Caucasian individuals) [5]. This compares with a probability of identity of 0.034 for the first variable region, 0.047 for the second variable region and 0.014 for the 2 regions taken together for the database described in this paper.

Although the majority of sequences were only represented once in the database, one common genotype was observed in the 4 samples, 4, 58, 78 and 92 (Fig. 1), which was very similar to the consensus sequence with the exception of a gap at base 302.3 (found in 41% of the sequences).

Stoneking et al. [5] and Orrego and King [1] determined the average number of pairwise differences between sequences where the total number of comparisons was $[n \times (n-1)]/2$, in which n was the number of samples. Stoneking et al. assumed selective neutrality and absence of population substructuring. Orrego and King as-

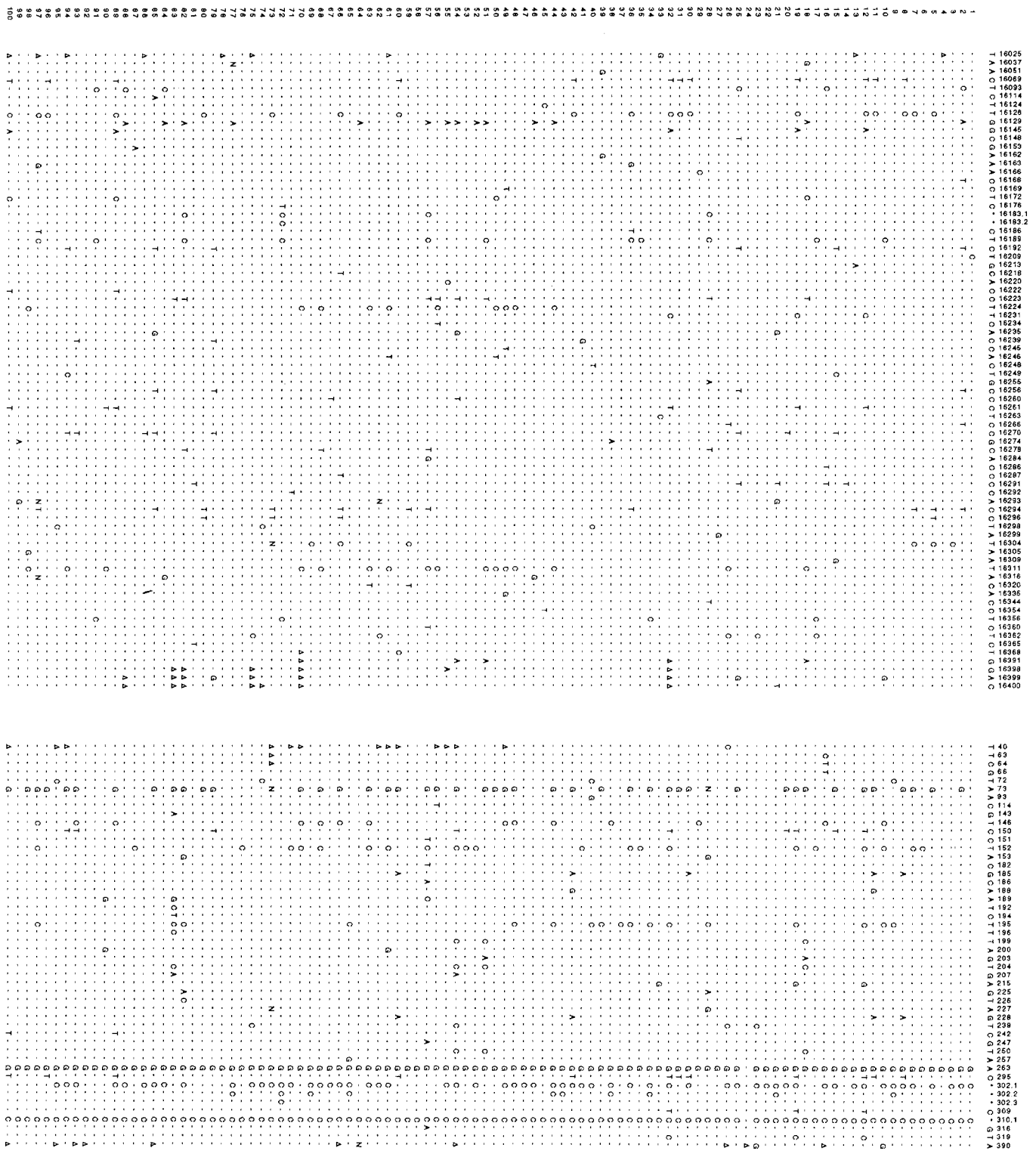

Fig. 1. List of 124 polymorphic nucleotide positions observed in sequences obtained from 100 unrelated white Caucasian samples, shown as differences from the Anderson reference sequence [32]. Nomenclature is in accordance with Anderson et al, and numbers followed by a decimal point indicate addition of nucleotides not found in this sequence. A dot (●) indicates sequence unchanged from the reference sequence, N indicates no nucleotide assignment and a triangle (△) indicates undetermined sequence. The relative frequencies of the most polymorphic sites identified in the paper of Stoneking et al. [5] were compared with our British White Caucasian data to determine whether their frequencies for Caucasians were indicative of the population as a whole. Many differences were observed between the two data sets including: the polymorphism at p16362 is at least 50% more frequent in Stoneking et al; the polymorphisms at p73 is 30% more frequent in Stoneking et al; a single addition at p302.3 is 25% more frequent in our data and the number of blanks (assuming these are due to double transitions or transversions) are much more frequent in Stoneking et al. (87 in 142 samples) than would be expected from our data (11 in 100 samples including positions of undetermined sequence)

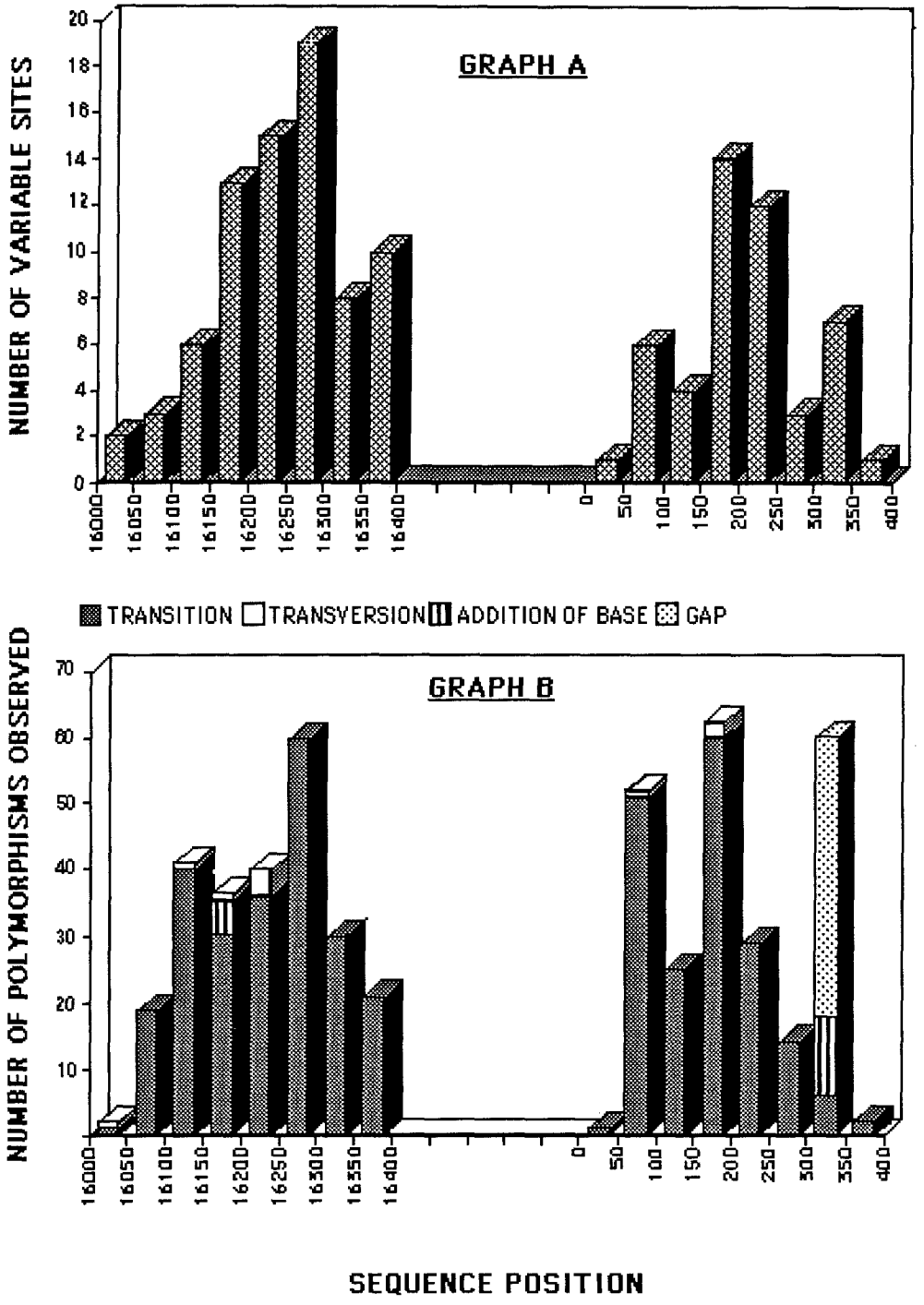

Fig. 2A B Graph showing **A**the number of variable sites and **B** the frequency of polymorphisms in blocks of 50 bases throughout the sequenced area of the non-coding region. Sequences obtained from 100 unrelated white Caucasians were compared to a consensus reference sequence (see text). The types of variation observed indicate that each type of mutation is not randomly distributed throughout the area sequenced e.g. additions and gaps (the formation of which are consistent with the slippage model of frame shift mutagenesis) are confined to a block of 50 bases in both the first and second variable regions

sumed that the distribution of pairwise differences approximated a Poisson distribution (also suggested by Di Rienzo and Wilson [25]). An estimate of genetic identity, fitting data from the first variable segment to this distribution (14 samples, average number of pairwise differences between any pair of sequences = 5.9), was calculated as 0.003 [1]. Pairwise comparisons for our data gave values for the average number of pairwise differences of 4.59 for the first variable region, 3.89 for the second region and 8.48 for the 2 regions taken together. However, our data does not approximate Poisson ($P << 0.001$) (Fig. 3A), probably because of the excess number of transitions present and the unequal substitution rate at each nucleotide position (Fig. 2).

It has been suggested [26, 33] that pairwise comparisons of mtDNA data may approximate a negative binomial distribution, which does not assume that each position within the region sequenced has an equal apportunity of change. However, when our data was compared to the negative binomial distribution (Fig. 3B), the fit was greatly improved but still significantly different ($0.01 < P < 0.025$) possibly because of the excess number of transitions present. Other statistical models are currently under investigation.

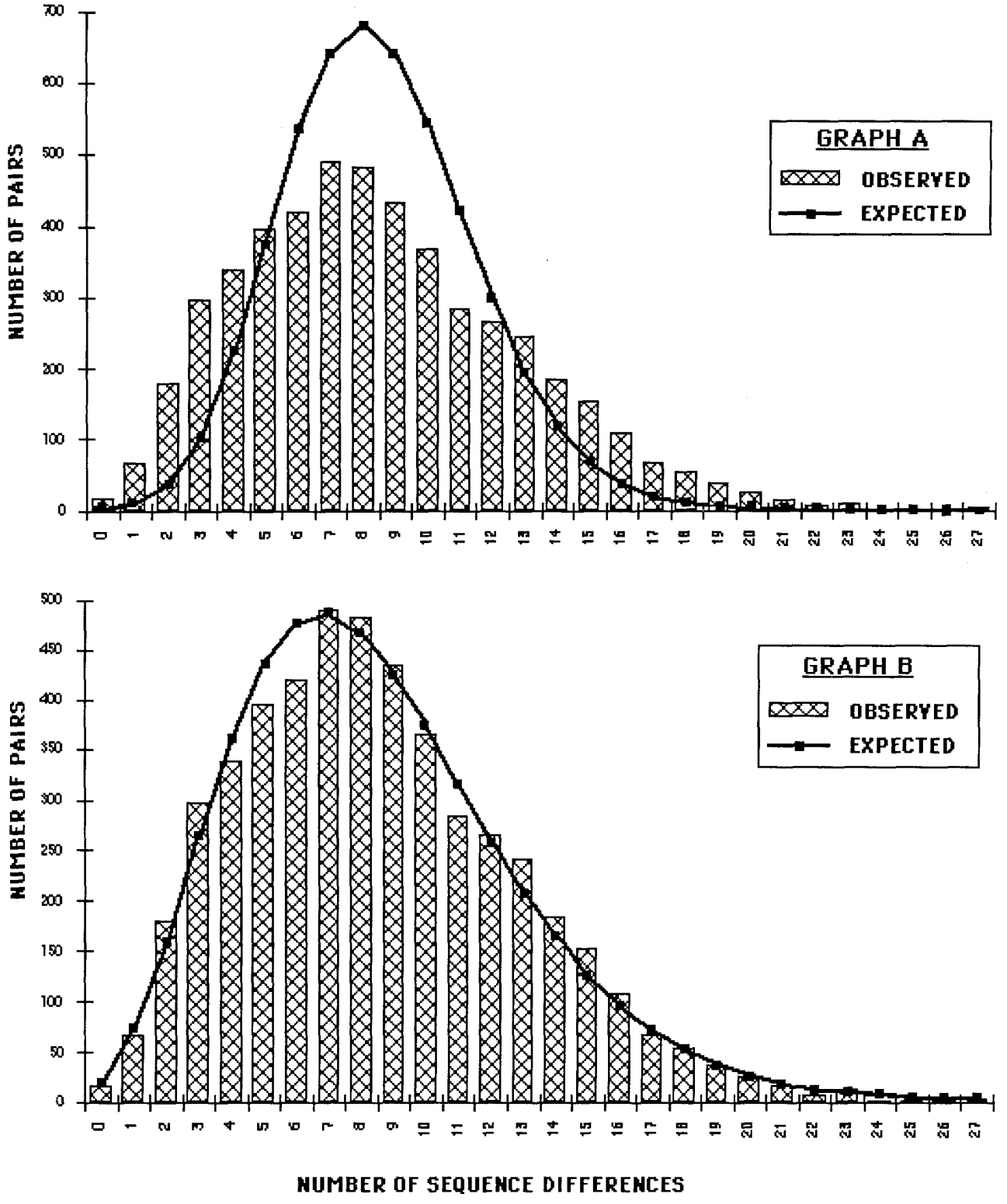

Fig. 3A, B. Graph showing that the distribution of pairwise differences for the sequenced data differs significantly from both **A** a Poisson distribution ($P << 0.001$) and **B** a negative binomial distribution ($0.01 < P < 0.025$)

Previous inter- and intra-population studies using mtDNA data have used phylogenetic analysis to establish either the maternal lineage of sequences and/or the origin of the human species [5, 7, 12–31]. These studies used information obtained from either the first of the variable regions or from the 2 regions taken together. Initial intrapopulation analysis of our data was carried out on the 2 variable regions separately using the PAUP (Phylogenetic Analysis using Parsimony) computer programme [34]. This analysis showed that the few sequences that grouped together in the first of the variable regions did not usually cluster in the second (unpublished data).

Further phylogenetic analysis may be required, but these preliminary results suggested that conclusions about individual subpopulation groups based on analysis of one region only or both regions together are problematical. Interpopulation studies using systems such as PAUP may be more informative.

From our analysis of white Caucasian mtDNA sequences we have shown that to obtain maximum information both of the variable regions present within the non-coding region must be sequenced. Further analysis is needed to determine the best fit to a statistical distribution, which could be useful when calculating probability of identity. We have also suggested that intrapopulation investigations based on the phylogeny of sequences using a system such as PAUP may be problematical.

This investigation shows that although there is currently no suitable statistical model to analyse the data to its full potential, the technique can still be used as a highly informative forensic test especially for the investigation of difficult sample types.

References

1. Orrego C, King MC (1990) Determination of familial relationships. In: Innis MA, Gelfand DH, Sninsky JJ, White TJ (eds) PCR protocols: A guide to methods and applications. Academic press, London and San Diego, pp 416–426
2. Hopgood R, Sullivan KM, Gill P (1992) Strategies for automated sequencing of human mitochondrial DNA directly from PCR products. Biotechniques 13:82–92
3. Sullivan KM, Hopgood R, Gill P (1992) Identification of human remains by amplification and automated sequencing of mitochondrial DNA. Int J Leg Med 105:83–86
4. Sullivan KM, Hopgood R, Lang B, Gill P (1991) Automated amplification and sequencing of human mitochondrial DNA. Electrophoresis 12:17–21
5. Stoneking M, Hedgecock D, Higuchi RG, Vigilant L, Erlich HA (1991) Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence specific oligonucleotide probes. Am J Hum Genet 48: 370–382
6. Ginther C, Issel-Tarver L, King M (1992) Identifying individuals by sequencing mitochondrial DNA from teeth. Nat Genet 2:135–138
7. Wrischnik LA, Higuchi RG, Stoneking M, Erlich HA, Arnheim N, Wilson AC (1987) Length mutations in human mtDNA: direct sequencing of enzymatically amplified DNA. Nucleic Acids Res 15:529–542
8. Hess JF, Parisi MA, Bennett JL, Clayton DA (1991) Impairment of mitochondrial transcription termination by a point mutation associated with the MELAS subgroup of mitochondrial encephalmoyopathies. Nature 351:236–239
9. McShane MA, Hammans SR, Sweeney M, Holt IJ, Beattie TJ, Brett EM, Harding AE (1991) Pearson Syndrome and mitochondrial encephalomyopathy in a patient with a deletion of mtDNA. Am J Hum Genet 48:39–42
10. Lestienne P (1992) Mitochondrial DNA mutations in human diseases: a review. Biochimie 74:123–130
11. Trischler H-J, Andreetta F, Moreas CT, Bonilla E, Arnaudo E, Danon MJ, Glass S, Zelaya BM, Vamos E, Telerman-Toppet N, Shanske S, Kadenbach B, DiMauro S, Schon EA (1992) Mitochondrial myopathy of childhood associated with depletion of mitochondrial DNA. Neurology 42:209–217
12. Aquadro CF, Greenberg BD (1983) Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals. Genetics 103:287–312
13. Kocher TD, Wilson AC (1991) Sequence evolution of mitochondrial DNA in humans and chimpanzees: control region and a protein-coding regoin. In: Osawa S, Honjo T (eds) Evolution of life. Springer, Tokyo. pp 391–413
14. Lundstrom R, Tavare S, Ward RH (1992) Estimating substitution rates from molecular data using the coalescent. Proc Natl Acid Sci USA 89:5961–5965
15. Hasegawa M, Horai S (1991) Time of the deepest root for polymorphism in human mitochondrial DNA. J Mol Evol 32: 37–42
16. Pesole G, Sbisa E, Preparata G, Saccone C (1992) The evolution of the mitochondrial D-loop region and the origin of modern man. Mol Biol Evol 9:587–598
17. Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. Science 253:1503–1507
18. Thomas WK, Beckenbach AT (1989) Variation on Salmonid mitochondrial DNA: evolutionary constraints and mechanisms of substitution. J Mol Evol 29:233–245
19. Horei S, Matsunaga E (1986) MtDNA polymorphism in Japanese. Hum Genet 72:105–117
20. Merriwether DA, Clark AG, Ballinger SW, Schurr TG, Soodyall H, Jenkins T, Sherry ST, Wallace DC (1991) The structure of human mitochondrial DNA variation. J Mol Evol 33:543–555
21. Brega A, Gardella R, Semino S, Morpurgo G, Astaldi Ricotti GB, Wallace DC, Santachiara Benerecetti AS (1986) Genetic studies on the Tharu population of Nepal: restriction endonuclease polymorphisms of mtDNA. Am J Hum Genet 39:502–512
22. Denaro M, Blanc H, Johnson MJ, Chen KH, Wilmsen E, Cavalli-Sforza LL, Wallace DC (1981) Ethnic variation in Hpa I endonuclease cleavage patterns of human mitochondrial DNA. Proc Natl Acad Sci USA 78:5768–5772
23. Bonné-Tamir B, Johnson MJ, Natali A, Wallace DC, Cavalli-Sforza LL (1986) Human mtDNA types in two Israeli populations – a comparative study at the DNA level. Am J Hum Genet 38:341–351
24. Vigilant L, Pennington R, Harpending H, Kocher TD (1989) Mitochondrial DNA sequences in single hairs from a Southern African population. Proc Natl Acad Sci USA 86:9350–9354
25. Di Rienzo A, Wilson AC (1991) Branching pattern in the evolutionary tree for human mitochondrial DNA. Proc Natl Acad Sci USA 88:1597–1601
26. Irwin DM, Kocher TD, Wilson AC (1991) Evolution of the cytochrome b gene of mammals. J Mol Evol 32:128–144
27. Horai S, Kondo R, Murayama K, Hayashi S, Koike H, Nakai N (1991) Phylogenetic affiliation of ancient and contemporary humans inferred from mitochondrial DNA. Philos Trans R Soc Lond [Biol] 33:409–417
28. Maddison DR (1991) African origin of human mitochondrial DNA reexamined. Syst Zool 40:355–363
29. Templeton AR, Hedges SB, Kumar S, Tamura K, Stoneking M (1991) Human origins and analysis of mitochondrial DNA sequences. Science 255:737
30. Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325:457–465
31. Nei M (1982) Evolution of human races at the gene level. In: Human genetics part A: The unfolding genome. Alan R. Liss, New York, pp 167–181
32. Anderson S, Bankier AT, Barrell BG, de Brujin MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJH, Staden R, Young IG (1981) Sequence and organisation of the human mitochondrial genome. Nature 290:457–465
33. Greenberg BD, Newbold JE, Sugino A (1983) Intraspecific nucleotide sequence variability surrounding the origin of replication in human mitochondrial DNA. Gene 21:33–49
34. Swofford DL (1989) PAUP, Phylogenetic Analysis Using Parsimony. Illinois Natural History Survey, Champaign, USA. Version 3.0 g