Am. J. Hum. Genet. 48:370–382, 1991

Population Variation of Human mtDNA Control Region Sequences Detected by Enzymatic Amplification and Sequence-specific Oligonucleotide Probes

Mark Stoneking,* Dennis Hedgecock,† Russell G. Higuchi,* Linda Vigilant,‡,[1] and Henry A. Erlich*

*Department of Human Genetics, Cetus Corporation, Emeryville, CA; †Bodega Marine Laboratory, University of California, Bodega Bay, CA; and ‡Division of Biochemistry and Molecular Biology, University of California, Berkeley

Summary

A method for detecting sequence variation of hypervariable segments of the mtDNA control region was developed. The technique uses hybridization of sequence-specific oligonucleotide (SSO) probes to DNA sequences that have been amplified by PCR. The nucleotide sequences of the two hypervariable segments of the mtDNA control region from 52 individuals were determined; these sequences were then used to define nine regions suitable for SSO typing. A total of 23 SSO probes were used to detect sequence variants at these nine regions in 525 individuals from five ethnic groups (African, Asian, Caucasian, Japanese, and Mexican). The SSO typing revealed an enormous amount of variability, with 274 mtDNA types observed among these 525 individuals and with diversity values, for each population, exceeding .95. For each of the nine mtDNA regions significant differences in the frequencies of sequence variants were observed between these five populations. The mtDNA SSO-typing system was successfully applied to a case involving individual identification of skeletal remains; the probability of a random match was approximately 0.7%. The potential useful applications of this mtDNA SSO-typing system thus include the analysis of individual identity as well as population genetic studies.

Introduction

A powerful new technique for the detection of molecular genetic variation in human populations is the analysis of DNA sequences that have been amplified by PCR and then hybridized with sequence-specific oligonucleotide (SSO) probes (Saiki et al. 1986; Goedde et al. 1989; Helmuth et al. 1990). In the present study we present the design and implementation of an SSO-typing system for variation in human mtDNA control-region sequences.

The control region is the major noncoding portion of the human mtDNA genome and includes the origin of replication of one strand, the D-loop region, and both origins of transcription (Anderson et al. 1981). We chose to analyze the control region because it is the most polymorphic region of the human mtDNA genome (Aquadro and Greenberg 1983; Cann et al. 1984, 1987; Horai and Hayasaka 1990), with most of the variation distributed not at random but rather concentrated in two hypervariable segments (Vigilant et al. 1989).

Here we present the nucleotide sequences of the two hypervariable segments from 52 individuals. These sequences are used to define nine candidate regions that appear to be suitable for SSO typing. We report the individual variant and mtDNA-type frequencies defined by this SSO-typing system in Caucasian, African, Asian, Japanese, and Mexican populations. SSO typing of PCR-amplified mtDNA control-region sequences reveals an enormous amount of diversity within and between these populations, attesting to the utility of this system for population genetic studies. We also demonstrate the power of the mtDNA SSO-

Received June 13, 1990; revision received September 17, 1990.

Address for correspondence and reprints: (Present address) Mark Stoneking. Department of Anthropology, Pennsylvania State University, University Park, PA 16802.

1. Present address: Department of Anthropology, Pennsylvania State University, University Park, PA 16802.

0002-9297/91/4802-0022$02.00

typing system for individual identification by applying it to a case involving identification of skeletal remains.

Subjects and Methods

Subjects

A total of 525 individuals from five ethnic groups (populations) were analyzed. All samples consisted of purified genomic DNA, with the exception of some sequences that were obtained from purified mtDNA. The five populations are as follows:

Caucasian.—There were 81 samples from CEPH (chosen from maternally unrelated individuals), 52 samples provided by E. Blake (Forensic Sciences Associated, Richmond, CA) from case studies, five samples from members of the laboratory of A. C. Wilson (University of California, Berkeley), and four published sequences (Anderson et al. 1981; Walberg and Clayton 1981; Aquadro and Greenberg 1983). Total number of samples was 142.

African.—There were 20 African-American samples from case studies provided by E. Blake, 99 African-American samples provided by M. C. King (University of California, Berkeley), seven Nigerian samples provided by J. Wainscoat (John Radcliffe Hospital, Oxford), and three published sequences (Aquadro and Greenberg 1983). Total number of samples was 129.

Asian.—There were 10 Chinese samples provided by R. Griffith (Cetus Corporation), 13 Southeast Asian samples provided by R. L. Cann (University of Hawaii, Honolulu), and 51 Southeast Asian samples from an alpha-thalassemia screening program, provided by S. Embury (University of California, San Francisco). Total number of samples was 74.

Japanese.—There were 86 samples from medical students participating in a study on responsiveness to a hepatitis vaccine, provided by T. Sasazuki (Kyushu University, Fukuoka, Japan).

Mexican.—There were 94 samples from Mexico City and environs, provided by C. Gorodezky (Instituto de Salubridad y Enfermedades Tropicales, Mexico).

PCR Amplification

A 1,024-bp portion of the mtDNA control region, encompassing the two hypervariable segments, was amplified using primers L15996 and H408 (fig. 1). Amplifications were performed in 100-μl volumes containing 20 pmoles of each primer, 0.1–0.5 μg of genomic DNA, 2 units of AmpliTaq™ DNA polymerase (Perkin Elmer–Cetus), and 50 μM concentration of each dNTP. Typically, 30 cycles of amplification were carried out in a programmable thermal cycler (Perkin Elmer–Cetus), with each cycle consisting of denaturation at 94°C for 45 s, annealing at 56°C for 1 min, and extension at 74°C for 1 min.

Asymmetric PCR and Sequencing

The two hypervariable segments of the control region were amplified separately using unequal primer ratios (e.g., asymmetric PCR) to generate single-stranded templates suitable for sequencing (Gyllensten and Erlich 1988). Segment I was amplified with primers L15996 and H16401, and segment II was amplified with primers L29 and H408 (fig. 1). Asymmetric PCR, purification of the PCR product by centrifugal filtration, and sequencing were performed according to methods described elsewhere (Vigilant et al. 1989).

Designing SSO Probes

The polymorphic nucleotides from 52 sequences (seven published previously and 45 reported here) of the two hypervariable segments of the mtDNA control

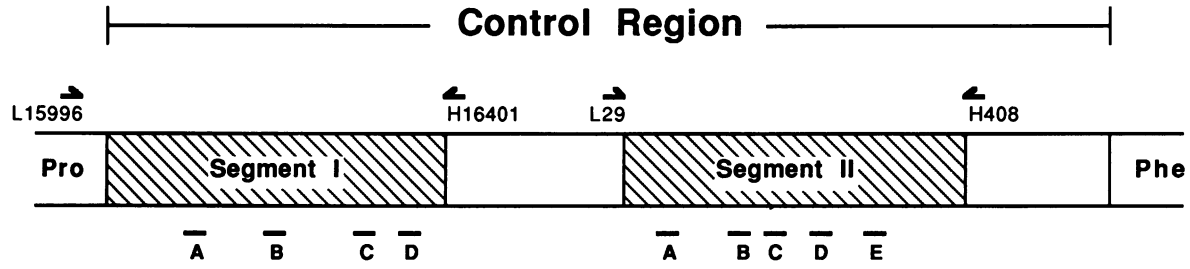
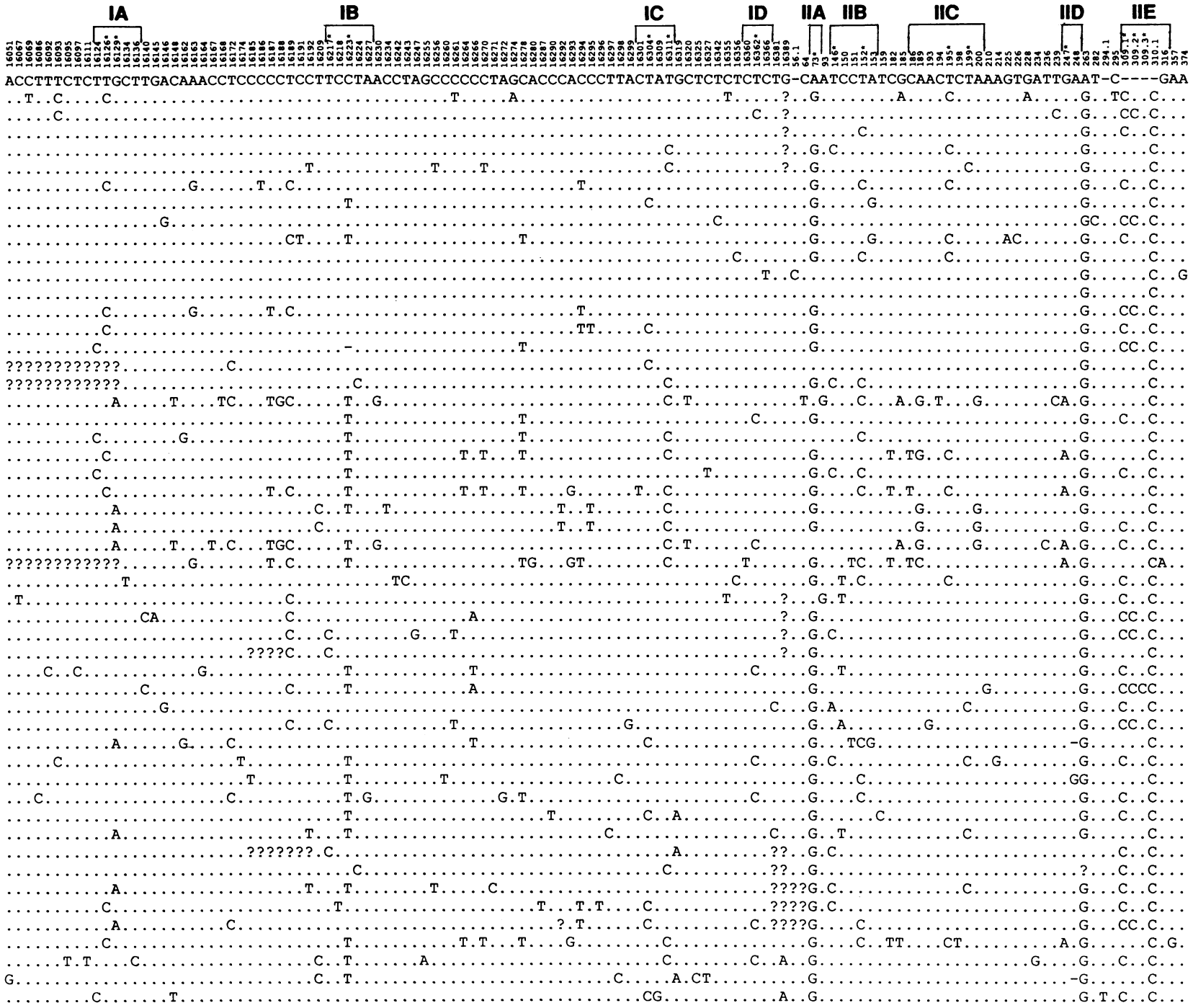


Figure 1 Diagram of human mtDNA control region. The control region is flanked by proline and phenylalanine tRNA genes. The location of the two hypervariable segments, the nine SSO-defined regions, and the primers used in PCR reactions are shown. PCR primer sequences are as given by Vigilant et al. (1989).


IA
IB
IC
ID
IIA
IIB
IIC
IID
IIE

region are shown in figure 2. In designing SSO probes, we looked for 15–20-bp regions in which most of the variation was confined to one or two polymorphic nucleotides. Visual inspection of these sequences revealed nine candidate regions (fig. 2), each with two or three major sequence variants, that would be potentially useful for SSO typing. The sequences of the 23 oligonucleotides that were designed to detect the major variants for these nine regions are given in table 1.

In designing these SSO probes, we made use of the fact that some mismatched basepairs are more destabilizing than others (Ikuta et al. 1987). In general, the following relationship holds: AT,GC>GT>GA,GG>AA,CA>TT,CT,CC, where ">" means "is more stable than." For example, the two strands of the mtDNA helix are designated as the L strand and the H strand (Anderson et al. 1981); if, at a particular nucleotide position, an A on the L strand (T on the H strand) is replaced by a G (C on the H strand), the appropriate SSO to detect the A variant would correspond to the L-strand sequence with an A at that position, while the appropriate SSO for the G variant would correspond to the H strand with a C at that position. Since most of the polymorphisms that occur in the mtDNA control region are transitions (fig. 2; Aquadro and Greenberg 1983; Vigilant et al. 1989), a set of rules can easily be devised for designing SSO probes. These rules are given in table 2.

The variants for eight of the regions consisted of nucleotide substitutions resulting in transitions, while for the ninth region (IIE) the variants consisted of changes in the length of a run of cytosine residues in the L strand. Initially, three SSO probes were designed to detect runs of cytosines of length 7, 8, and 9, respectively. While specific discrimination could be obtained for the shortest probe (IIE1), the IIE2 and IIE3 probes would not differentiate between runs of length 8 and runs of length 9 with any of the hybridization and wash conditions that we tried. We therefore used two SSO probes at this locus: IIE1, to detect a variant with a length of 7 cytosines, and IIE2, to detect a variant with a length of 8 or more cytosines.

Oligonucleotide Labeling

Oligonucleotides (table 1) were labeled with [γ-^{32}P]-ATP (Amersham) to a specific activity of at least 10^8 cpm/μg DNA. Unincorporated nucleotides were removed by centrifugal dialysis through a Centricon-10 microconcentrator (Amicon).

mtDNA Typing

Approximately 200 ng (5 μl) of each amplified DNA was added to denaturation buffer (0.4 N NaOH, 25 mM EDTA) and dotted on a nylon membrane (Biodyne). The DNA was fixed to the membrane by UV irradiation with a Stratalinker™ UV Crosslinker (Stratagene). Membranes were prehybridized in hybridization solution (5 × SSPE, 5 × Denhardt's, 0.5% SDS) for 30 min at the hybridization temperature (table 1). Labeled SSO probes were then added directly to the hybridization solution to a concentration of 1 pmol/ml, and hybridization was carried out for 2 h. Membranes were rinsed in wash solution (2 × SSPE, 0.1% SDS) at room temperature, washed at the hybridization temperature for 20 min, and exposed to film for 2–48 h.

Statistical Analysis

For each individual, the combination of sequence variants observed across all nine regions is referred to as an *mtDNA type*. An unbiased estimate of the genetic diversity (equivalent to heterozygosity) in each population, on the basis of mtDNA types, is

$$h = (1 - \Sigma x^2)\, n/(n-1)\,, \qquad (1)$$

where n is the sample size and where x is the frequency of each mtDNA type (Tajima 1989). The probability of two randomly selected individuals from a population having identical mtDNA types is simply

$$p = \Sigma x^2\,. \qquad (2)$$

For individuals from different populations, the probability of identity is the sum of the products of the

Figure 2 List of 122 polymorphic nucleotide positions and their state in 52 control-region mtDNA sequences, shown as differences from complete reference mtDNA sequence (Anderson et al. 1981). Nucleotide positions are numbered according to the method of Anderson et al. (1981), with numbers followed by a decimal point indicating additions of nucleotides not found in the reference sequence. Nucleotides encompassed by the nine SSO-defined regions are indicated, with asterisks (*) designating the polymorphic nucleotides detected by the SSO probes. A dash (–) indicates a deletion of a nucleotide, and a question mark (?) indicates undetermined sequence. Sequences 1–19 are from Caucasians (1 is from Anderson et al. [1981], 16 is from the KB cell line [Walberg and Clayton 1981], and 17 and 18 are from Aquadro and Greenberg [1983]); sequences 20–26 are from Nigerians; sequences 27–29 (Aquadro and Greenberg 1983) are from African-Americans; and the remaining sequences are from Asians.

Table 1

SSO Probes Used to Detect mtDNA Control-Region Sequence Variation

Region[a] and Variant	Polymorphic Nucleotides[b]	Hybridization Temperature (°C)	Probe Sequence (5′→3′)	Strand[c]
16118–16136:				
IA1	16126 T, 16129 G	55	ATGGTACCGTACAATATTC	H
IA2	16126 C, 16129 G	55	GAATATTGCACGGTACCAT	L
IA3	16126 T, 16129 A	50	GAATATTGTACAGTACCAT	L
16211–16228:				
IB1	16217 T, 16223 C	55	CAGCAATCAACCCTCAAC	L
IB2	16217 T, 16223 T	55	GTTGAAGGTTGATTGCTG	H
IB3	16217 C, 16223 C	55	CAGCAACCAACCCTCAAC	L
16300–16317:				
IC1	16304 T, 16311 T	45	TTATGTACTATGTACTGT	H
IC2	16304 C, 16311 T	45	ACAGCACATAGTACATAA	L
IC3	16304 T, 16311 C	45	ACAGTACATAGCACATAA	L
16357–16374:				
ID1	16362 T	55	TCATCCATGGGGACGAGA	H
ID2	16362 C	55	TCTCGCCCCCATGGATGA	L
68–83:				
IIA1	73 A	55	GGGGTATGCACGCGAT	L
IIA2	73 G	55	ATCGCGTGCACACCCC	H
141–158:				
IIB1	146 T, 152 T	55	AATAATAGGATGAGGCAG	H
IIB2	146 C, 152 T	50	CTGCCCCATCCTATTATT	L
IIB3	146 T, 152 C	50	CTGCCTCATCCCATTATT	L
186–204:				
IIC1	195 T, 199 T	55	ACTTTAGTAAGTATGTTCG	H
IIC2	195 C, 199 T	50	CGAACATACCTACTAAAGT	L
IIC3	195 T, 199 C	45	CGAACATACTTACCAAAGT	L
240–258:				
IID1	247 G	55	GTGCAGACATTCAATTGTT	H
IID2	247 A	55	AACAATTAAATGTCTGCAC	L
303–316:[d]				
IIE1	7-length cytosines	65	AAACCCCCCCTCCCCCCG	L
IIE2	8+-length cytosines	65	AAACCCCCCCCTCCCCCCG	L

[a] Nucleotides encompassed by the SSO probes, numbered according to the published reference sequence (Anderson et al. 1981).

[b] Nucleotide position and state on the L strand detected by the SSO probe.

[c] H-probe sequence corresponding to heavy strand of published sequence (Anderson et al. 1981); L-probe sequence corresponding to light strand.

[d] Variation for this region is in the length of a run of cytosines between nucleotides 303 and 309. At nucleotides 311–315 there is a run of five cytosines in the reference sequence; however, in all other individuals sequenced, this is a run of six cytosines (fig. 2), and the IIE1 and IIE2 probe sequences reflect this.

mtDNA type frequencies from each population. The probability of identity can also be estimated from the control-region nucleotide sequences. One way is to use equation (2) above and define mtDNA types by the nucleotide sequences; however, this method relies only on the frequency of each type and does not consider the amount of difference (i.e., number of substitutions) between types. Two other ways that do utilize the amount of difference between types rely on the result from neutral theory that the probability that two sequences will be identical is expected to be $1/(1+M)$ (Watterson 1975). For mitochondrial DNA, $M = N\mu$, where N is the effective population size and where μ is the neutral mutation rate. Tajima

Table 2

Rules for Designing mtDNA SSO Probes

Polymorphism[a]	Probe 1		Probe 2	
	Strand	Nucleotide	Strand	Nucleotide
T→C	H	A	L	C
C→T	L	C	H	A
G→A	H	C	L	A
A→G	L	A	H	C

Note.—The strand and nucleotide state at the variant position for the standard (Probe 1) and mutant (Probe 2) sequences are shown.

[a] Shown as changes from the L-strand sequence (Anderson et al. 1981).

(1989) gives equations for estimating M both from the number of polymorphic nucleotide positions and from the average number of nucleotide differences between any pair of sequences. These two separate estimates of M thus yield corresponding estimates of the probability of identity of two randomly chosen sequences that do take into account the amount of difference between sequences.

Results

SSO Typing of Known Sequences

SSO typing of mtDNA control-region sequence variation was performed on 12 individuals of known sequence, in order to verify the accuracy of the method. These 12 sequences were chosen so as to include all of the possible sequence variants at each of the nine SSO-defined mtDNA regions, including at least one expected "blank" variant. Blank variants are caused by nucleotide substitutions that prevent the formation of stable DNA heteroduplexes with *any* of the SSO probes for a particular region.

In each case in which the previously determined sequence predicted that a probe should hybridize, the expected results (specific hybridization with only the corresponding probe) were obtained. Of 16 instances in which the sequence predicted a mismatch with all of the SSO probes for a region, 14 typed as blanks as expected, while two hybridized to one of the probes (data not shown). For these latter two instances the mismatch occurred at the last position of the SSO probe, and these were the only instances in which this position was mismatched. For two of the 14 instances that typed as blanks as expected, the mismatch occurred one position in from the end of the probe. Hence, we conclude that mismatches at the extreme ends of the probe will not prevent hybridization but that all other mismatches (even those that are one position in from the end) will prevent hybridization of the SSO probes.

Absence of Cross-Hybridization

Judicious manipulation of the hybridization and wash conditions eliminated cross-hybridization between all but two of the 23 probes. Consistently significant cross-hybridization was observed with the IIE1 and IIE2 probes (e.g., see fig. 3). These probes were the only ones that detected variation in the length of a region, not nucleotide substitutions within a region. It is possible that the cross-hybridization reflects the presence of heteroplasmy, i.e., intraindividual variation for the two length variants detected by the probes. Alternatively, the apparent heteroplasmy could be generated in vitro during amplification via PCR. However, the most likely explanation is that the probes are truly cross-hybridizing to the alternative sequence variant, since direct sequencing of the PCR products generally yielded unambiguous, clean sequence ladders in the IIE region (data not shown).

Population Variation

The frequencies of the sequence variants (including blanks) for the nine mtDNA regions are given in table 3. Although the blanks in a population are heterogeneous mixtures of different nucleotide sequences, phenotypically they type as a single sequence variant and are therefore treated as such in subsequent analyses.

Polymorphism was found for every region in every population, with the exception of the IIA region, which was fixed for the IIA2 variant in the Japanese population. Blanks tended to be found at relatively low frequencies (less than 20%), except for the IIB

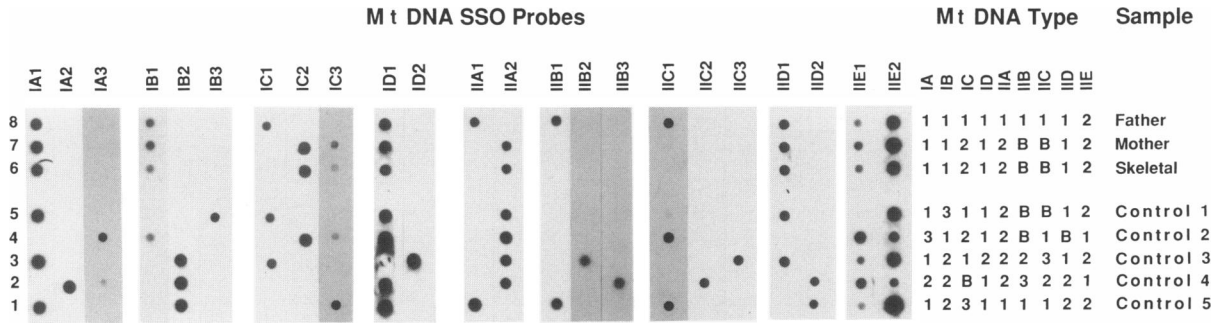


Figure 3 Results of case analysis involving mtDNA typing via SSO probes. Shown are the hybridization results from 23 probes, along with the resulting mtDNA types of five control samples of known sequence and of the father, mother, and skeletal sample from the case discussed in the text. The IC3 probe shows slight cross-hybridization with the IC2 sequence. There is rather more cross-hybridization evident between the IIE1 and IIE2 probes, probably because these probes detect variation in the length of a run of consecutive cytosine residues and not nucleotide substitutions, as discussed in the text. Despite the cross-hybridization, the mtDNA types can be unambiguously determined.

region in all populations and the IIC region in Africans.

The number of mtDNA types in a population ranged from 47 to 99, with 274 types observed in the total sample of 525 individuals (a list of the frequency of each type in each population can be obtained from the authors on request). Additional characteristics of population variation, based on mtDNA types, are given in table 4. The probability of two unrelated individuals would have identical mtDNA types ranges from 1.8% to 4.4%, and the genetic diversity values for each population are all more than .95. By all of these measures, then, there is an enormous amount of variation detected by SSO typing of mtDNA control-region sequences.

Population Differentiation

Each of the nine regions listed in table 3 displays significant differences in the frequencies of the sequence variants across the five populations, by the χ^2 test of heterogeneity (Sokal and Rohlf 1981). Particularly dramatic frequency differences are observed for the following: the IB1 variant, with a frequency of 76.0% in Caucasians and 3.2%–32.4% elsewhere; the IB3 variant, with a frequency of 22.3% in Mexicans and 1.4%–13.5% elsewhere; the IIA1 variant, with a frequency of 40.1% in Caucasians and 0.0%–8.5% elsewhere; the IIB1 variant, with a frequency of 27.9% in Africans and 39.4%–56.3% elsewhere; and the IICl variant, with a frequency of 35.7% in Africans and 67.6%–88.3% elsewhere.

The relative amount of mtDNA-type sharing within and between populations is further evidence of population differentiation. The probability of two unrelated individuals having identical mtDNA types is four to nine times higher if they come from the same population than if they come from different populations (table 4).

A Case Analysis

In order to demonstrate the utility of mtDNA typing for individual identification, we present the results of a case analysis. In October 1984, a 3-year-old child disappeared from her parent's home; in March 1986, skeletal remains of a human child were found in the desert 2 miles from the parents' residence. It was therefore of interest to determine whether the skeletal remains could have come from the missing child.

The results of SSO typing of PCR-amplified mtDNA control-region sequences from the parents and from the skeletal sample are shown in figure 3. The skeletal sample and the mother have identical mtDNA types and differ from the father at four of the nine regions. The mtDNA type of the skeletal sample is therefore consistent with that of a child of the mother. This mtDNA type was not observed previously in the sample of 142 Caucasians; hence we estimate that the probability that an unrelated individual would have this mtDNA type is 1/143 = 0.7%.

In order to confirm this result the sequences of the two hypervariable mtDNA segments were obtained from the parents and from the skeletal sample. The mother and the skeletal sample had identical sequences across both segments, with a total of nine

Table 3

Frequencies of Sequence Variants at Nine mtDNA SSO-defined Regions for Five Populations

Region and Variant	Population Frequency (%)					
	Caucasian ($n = 142$)	African ($n = 129$)	Asian ($n = 74$)	Japanese ($n = 86$)	Mexican ($n = 94$)	Total ($n = 525$)
IA:						
1	69.7	55.8	68.9	80.2	88.3	71.2
2	18.3	8.5	4.1	2.3	1.1	8.2
3	8.5	17.8	22.9	17.4	9.5	14.5
Blank	3.5	17.8	4.1	.0	1.1	6.1
IB:						
1	76.0	10.1	32.4	11.6	3.2	30.1
2	10.6	77.5	44.6	75.6	67.0	52.6
3	1.4	4.7	13.5	5.8	22.3	8.4
Blank	12.0	7.7	9.5	7.0	7.4	9.0
IC:						
1	74.6	58.1	58.1	81.4	83.0	70.9
2	9.9	.8	17.6	7.0	1.1	6.7
3	12.7	27.1	18.9	9.3	12.7	16.5
Blank	2.8	14.0	5.4	2.3	3.2	5.9
ID:						
1	86.6	69.8	77.0	53.5	50.0	69.1
2	9.9	18.6	23.0	46.5	47.9	26.7
Blank	3.5	11.6	.0	.0	2.1	4.2
IIA:						
1	40.1	3.9	2.7	.0	8.5	13.7
2	59.9	95.3	97.3	100.0	91.5	86.1
Blank	.0	.8	.0	.0	.0	.2
IIB:						
1	56.3	27.9	48.6	51.2	39.4	44.4
2	8.5	1.6	9.5	7.0	10.6	7.0
3	15.5	11.6	8.1	17.4	7.5	12.4
Blank	19.7	58.9	33.8	24.4	42.5	36.2
IIC:						
1	67.6	35.7	74.3	73.3	88.3	65.3
2	17.6	19.4	4.0	3.5	4.3	11.4
3	2.8	1.5	14.9	8.1	2.1	5.0
Blank	12.0	43.4	6.8	15.1	5.3	18.3
IID:						
1	95.1	74.4	82.4	83.7	79.8	83.6
2	1.4	19.4	1.4	.0	1.1	5.5
Blank	3.5	6.2	16.2	16.3	19.1	10.9
IIE:						
1	38.0	45.7	28.4	43.0	28.7	37.7
2	57.8	44.1	71.6	53.5	66.0	57.2
Blank	4.2	10.1	.0	3.5	5.3	5.1

shared differences from the published reference sequence (data not shown). Sequencing confirmed the mtDNA SSO-typing results shown in figure 3, revealing that the substitutions accounting for the blanks in both the mother and the skeletal sample at the IIB and IIC regions were C→T at nucleotide 150 and A→G at nucleotide 189, respectively.

Discussion

The use of a dot-blot typing system based on PCR-amplified DNA is a simple, rapid, and powerful approach for analyzing mtDNA variation in human populations. Use of the PCR makes minute quantities of DNA, such as might be found in single hairs (Higuchi

Table 4

Parameters of Population Variation, Based on SSO Typing of Human mtDNA Control-Region Sequences

Population (*n*)	No. of Types	Frequency of Most Common Type (%)	Diversity[a]	Probability of Identity[b] within Populations (%)	Probability of Identity between Populations[c] (%)	Ratio[d]
Caucasian (142)	99	6.3	.98	1.9	.2	9.5
African (129)	82	5.4	.98	1.8	.4	4.5
Asian (74)	57	6.8	.98	2.3	.5	4.6
Japanese (86)	58	8.1	.97	2.6	.5	5.2
Mexican (94)	47	11.7	.96	4.4	.7	6.3
Total (525)	274	3.4	.99	—	—	—
Average	—	—	—	2.6	.5	5.2

[a] Calculated according to equation (1).
[b] Calculated according to equation (2).
[c] Average between each population and the other four populations.
[d] Probability of identity within populations/probability of identity between populations.

et al. 1988) or in highly degraded forensic samples (von Beroldingen et al. 1989), amenable to analysis. The amount of diversity detected by SSO typing of mtDNA control-region sequences is enormous, with genetic diversity values exceeding .95 in each population (table 4). This is greater than the amount of heterozygosity (comparable to diversity) detected in highly polymorphic loci in nuclear DNA, such as the HLA DQα locus (Helmuth et al. 1990) or VNTR loci (Nakamura et al. 1987). An advantage of the dot-blot typing system over the gel-based assays for VNTR or DNA minisatellite variation is that the determination of a match between two samples is not confounded by gel artifacts such as band shifting (Lander 1989). Moreover, uncertainties arise in determining the probability of a random match by VNTR or DNA minisatellite typings (Cohen 1990), since the assumption of Hardy-Weinberg equilibrium required to compute genotype frequencies from observed allele frequencies is often violated by VNTR systems (Lander 1989; Helmuth et al. 1990). For the SSO-defined mtDNA types, the probability of a random match is directly calculated from the mtDNA-type frequencies and hence does not rely on any additional assumptions such as Hardy-Weinberg equilibrium.

Undetected Variation

However, a potential disadvantage of the SSO-typing system concerns undetected variation. Undetected variation can arise in a number of ways. The first way, mentioned previously, is that blanks occur at each region. Two individuals that type as blanks need not have identical nucleotide sequences for that region. We have attempted to design the SSO probes to minimize the occurrence of blanks and have largely been successful in that the frequency of blanks for each region is less than 20%, with the exception of the IIB region in all populations and the IIC region in Africans (table 3). Additional sequencing of blank variants at these regions has revealed candidate SSO probe sequences that would potentially reduce the frequency of blanks to less than 20% (M. Stoneking, unpublished results).

Another way in which variation might remain undetected is that some nucleotide substitutions may not be sufficiently destabilizing to prevent hybridization of a mismatched SSO probe. Two individuals might thus type as identical for a region, even though they have different nucleotide sequences. The SSO typing of a number of individuals of known sequence revealed that mismatches at the extreme end of a probe sequence will not prevent hybridization but that all other mismatches are sufficient to prevent hybridization of an SSO probe. Thus, there is a strong inference that two individuals that type as positive for an SSO probe do in fact have identical nucleotide sequences over the region detected by the SSO probe.

The last way in which variation might go undetected is that, since the SSO-defined regions constitute only a portion of the mtDNA control-region sequence, there will be undetected substitutions at positions outside these regions. Two individuals who have identical mtDNA types, as determined by SSO typing, might

not have identical control-region sequences. We investigated this by determining the nucleotide sequence of the two hypervariable segments of the control region for nine Caucasians who had identical SSO-defined mtDNA types. Four different sequences were found among these nine individuals (data not shown), with one sequence shared by five individuals, another sequence shared by two individuals, and two sequences each represented by a single individual. This limited survey indicates that there is a greater-than 50% probability that individuals with identical SSO-defined mtDNA types do in fact possess identical control-region sequences. As discussed further below, for some applications (e.g., population genetic studies) this level of resolution is probably adequate, while for other applications (e.g., individual identification) it would be desirable to determine the control-region nucleotide sequences of samples with identical SSO-defined mtDNA types.

Applications for Population Genetic Studies

SSO typing of mtDNA control-region sequences that have been amplified by PCR should be a valuable approach for human population genetic studies. Large numbers of samples can be readily typed on the basis of easily obtained biological materials such as plucked hairs (Vigilant et al. 1989), making it feasible to study remote, isolated populations that are of anthropological interest. Furthermore, the level of discrimination between individuals and among populations is almost as great as can be achieved with complete sequences of the hypervariable segments. The relative power of SSO typing in detecting variation can be assessed by comparing the mtDNA types that would be inferred by SSO typing of the 52 sequences shown in figure 2 to the actual sequence-defined mtDNA types. SSO typing would detect 49 mtDNA types with an overall diversity of .997. The actual sequences reveal that there are 52 types (e.g., every individual has a distinct sequence), for an overall diversity of 1.0. Thus, SSO typing reveals nearly as much diversity as is revealed by the complete nucleotide sequences of the hypervariable control-region segments, at a fraction of the time, effort, and cost of obtaining such sequences.

The five populations in the present study differed significantly in the frequencies of the sequence variants at each of the nine SSO-defined regions. However, in other populations this might not be the case: some of these regions might not be informative or polymorphic, and there may be other regions and/or sequence variants that would provide more information. We therefore recommend that the nucleotide sequences of 25–50 individuals should be obtained from any new population that is to be studied. These sequences can then be analyzed to define appropriate regions for SSO typing of the remaining individuals. As the SSO typing proceeds, additional sequencing of regions with unacceptably high levels of blanks can further increase the amount of information obtained.

Applications for Individual Identification

There are several reasons why SSO typing of mtDNA control-region sequence variation should prove useful in ascertaining individual identification. First, mtDNA is present in high copy number, with an average of several hundred mtDNA molecules in each cell (Robin and Wong 1988). Thus, in those biological samples where there are minute amounts of DNA that may be highly degraded, there is a greater likelihood of success in analyzing mtDNA relative to single-copy nuclear genes, simply because of the higher copy number of mtDNA.

Second, human mtDNA is apparently strictly maternally inherited with no recombination (Giles et al. 1980), so every individual is haploid, possessing a single mtDNA type. The detection and analysis of mixed samples (i.e., samples containing DNA from two or more individuals) is thus considerably simplified. Also, all relatives who share the maternal lineage of an individual should possess identical mtDNA types, making it easier to infer a biological relationship on the basis of the mtDNA type of a sample.

Finally, mtDNA evolves rapidly, and the control region in particular shows an extremely high level of polymorphism in humans (Aquadro and Greenberg 1983; Cann et al. 1984, 1987; Vigilant et al. 1989; Horai and Hayasaka 1990). The results of this study bear this out. There are potentially 82, 944 different mtDNA types detected by the 23 SSO probes for the nine regions; in actuality 274 mtDNA types were detected in the present survey of 525 individuals. The average probability that two unrelated individuals would have identical mtDNA types is about 2.6% (from table 4). There is thus a correspondingly high probability—approximately 97.4%—of excluding individuals who truly did not contribute a particular biological sample. In practice the actual probability of exclusion will depend on the particular mtDNA type of the sample, but, since the frequency of the most common mtDNA type in each population ranges from 5.4% to 11.7% (table 4), the probability of exclusion will always be at least 88.3%–94.6%.

The identity probability based on SSO typing can be compared with that based on complete nucleotide sequences of the two hypervariable control-region segments. For the 19 Caucasian nucleotide sequences in figure 2, the probability of identity can be estimated in three different ways, as described above in the Subjects and Methods section. The first way is to simply use the sum of the square of the sequence-defined mtDNA-type frequencies, analogous to the computation for the SSO-defined mtDNA types (table 4); since the 19 Caucasian nucleotide sequences were all distinct, the resulting estimate of the probability of identity is 5.3%.

The other two ways take into consideration the amount of difference between types, as measured either by the number of polymorphic nucleotide sites or by the average number of pairwise nucleotide differences. When the equations given by Tajima (1989) are used, for polymorphic nucleotide sites $M = 16.59$ and the probability of identity is 5.7%, while for the average number of pairwise nucleotide differences $M = 16.44$ and the probability of identity is 5.5%.

It should be noted that these estimates have unknown statistical properties and, in particular, may behave quite differently when various assumptions involving neutrality—e.g., lack of population substructure and population equilibrium—are violated, as is undoubtedly the case for human populations (Whitman et al. 1986). In particular, the difference between the M estimate from the number of polymorphic nucleotide positions and that from the average number of pairwise nucleotide differences is the basis of a test for selection (Tajima 1989). Nevertheless, these three different ways of estimating the probability of identity from nucleotide sequences all gave approximately the same value, about 5.5%.

This value is much larger than a previous estimate of the probability of identity, i.e., 0.27%, that was based on partial mtDNA control-region sequences of 14 Caucasians (Orrego and King 1990). This latter value arises from fitting a Poisson curve to the distribution of the number of nucleotide sequence differences between all pairs of individuals. However, this procedure assumes that the observations are statistically independent, which is not the case for the set of all pairwise comparisons of individuals. The effect of violating this assumption has an unknown effect on the estimate of the probability of identity; since lack of statistical independence is not a problem with the methods we used, our estimate is preferable.

The sequence-based estimate of the probability of identity is about 5.5% for Caucasians. This is greater than the 1.9% value from the SSO typing, no doubt because the sample size ($n = 19$) for the nucleotide sequences is much smaller than the sample size ($n = 142$) for the SSO typing. Obviously, for the purpose of individual identification, SSO typing is not superior to nucleotide sequencing. Rather, the most powerful use of the mtDNA SSO-typing system for individual identification will arise from a combination of SSO typing with nucleotide sequencing. The SSO-typing system can provide a rapid determination of the mtDNA type, with a high probability of excluding random samples. Nucleotide sequences of the control-region hypervariable segments can then be determined for samples that are not excluded by the SSO typing, which (if identical) will provide further evidence of individual identification.

The sample case presented here, which followed the above procedure, attests to the utility, for individual identification, of SSO typing of PCR-amplified mtDNA control-region sequences. The mtDNA type of the skeletal sample matched the mtDNA type of the mother of the missing child; the probability that an unrelated individual would have the same mtDNA type was estimated to be 0.7%. In principle, VNTR loci (Balazs et al. 1989) or DNA minisatellites (Jeffreys et al. 1985) might give even greater assurance that the skeletal sample did indeed come from the missing child. However, it is by no means certain that an accurate VNTR or minisatellite typing could result from the minute quantities of highly degraded DNA that were obtained from the skeletal sample.

We anticipate that the mtDNA SSO-typing system will prove valuable not only in linking biological remains to missing individuals (as in the sample case above) but also in the analysis of material from sexual assault cases. There are approximately 50–100 mtDNA molecules in the midpiece of the sperm (Hecht et al. 1984), compared with about 500–1,000 mtDNA molecules in the average epithelial cell (Robin and Wong 1988). We have verified that SSO typing of the PCR-amplified mtDNA from a 10:1 mixture of sperm and epithelial cells reveals approximately equal signals from the sperm mtDNA type and the epithelial cell mtDNA type (M. Stoneking, unpublished results) and that the sperm mtDNA type can still be readily detected from an initial 1:1 mixture of sperm and epithelial cells. Since mtDNA is haploid, differences between the sperm mtDNA type of the perpetrator and the epithelial cell mtDNA type of the victim will be readily apparent on SSO typing of PCR-amplified mtDNA control-region sequences from the sexual as-

sault material. Subtracting the victim's mtDNA type, determined from a separate sample contributed by the victim, then allows a determination of the mtDNA type of the perpetrator.

The combination of enzymatic amplification via PCR with the SSO hybridization assay for detecting variation at the hypervariable mtDNA control-region sequences thus provides a powerful new tool for individual identification. Conversion of this typing system to a nonradioactive "reverse dot-blot format," in which the SSO probes are immobilized on a membrane (Saiki et al. 1989), is currently in progress and promises to make it even easier and faster to obtain results with this system.

Acknowledgments

We thank E. Blake, R. L. Cann, S. Embury, C. Gorodezky, R. Griffith, M. C. King, T. Sasazuki, and J. Wainscoat for providing DNA samples; C. Levenson and D. Spasic for oligonucleotide synthesis; P. Shabe for assistance with data analysis; and A. C. Wilson for valuable discussion.

References

Anderson S, Bankier AT, Barrell BG, de Bruijn MHL, Coulson AR, Drouin J, Eperon IC, et al (1981) Sequence and organization of the human mitochondrial genome. Nature 290:457–465

Aquadro CF, Greenberg BD (1983) Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals. Genetics 103:287–312

Balazs I, Baird M, Clyne M, Meade E (1989) Human population genetic studies of five hypervariable DNA loci. Am J Hum Genet 44:182–190

Cann RL, Brown WM, Wilson AC (1984) Polymorphic sites and the mechanism of evolution in human mitochondrial DNA. Genetics 106:479–499

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325:31–36

Cohen JE (1990) DNA fingerprinting for forensic identification: potential effects on data interpretation of subpopulation heterogeneity and band number variability. Am J Hum Genet 46:358–368

Giles RE, Blanc H, Cann HM, Wallace DC (1980) Maternal inheritance of human mitochondrial DNA. Proc Natl Acad Sci USA 77:6715–6719

Goedde HW, Singh S, Agarwal DP, Fritze G, Stapel K, Paik YK (1989) Genotyping of mitochondrial aldehyde dehydrogenase in blood samples using allele-specific oligonucleotides: comparison with phenotyping in hair roots. Hum Genet 81:305–307

Gyllensten UB, Erlich HA (1988) Generation of single-stranded DNA by the polymerase chain reaction and its application to direct sequencing of the HLA-DQα locus. Proc Natl Acad Sci USA 85:7652–7656

Hecht NB, Liem H, Kleene KC, Distel RJ, Ho SM (1984) Maternal inheritance of the mouse mitochondrial genome is not mediated by a loss or gross alteration of the paternal mitochondrial DNA or by methylation of the oocyte mitochondrial DNA. Dev Biol 102:452–461

Helmuth R, Fildes N, Blake E, Luce MC, Chimera J, Madej R, Gorodezky C, et al (1990) HLA-DQα allele and genotype frequencies in various human populations determined by using enzymatic amplification and oligonucleotide probes. Am J Hum Genet 47:515–523

Higuchi R, von Beroldingen CH, Sensabaugh GF, Erlich HA (1988) DNA typing from single hairs. Nature 332: 543–546

Horai S, Hayasaka K (1990) Intraspecific nucleotide sequence differences in the major noncoding region of human mitochondrial DNA. Am J Hum Genet 46:828–842

Ikuta S, Takagi K, Wallace RB, Itakura K (1987) Dissociation kinetics of 19 base paired oligonucleotide-DNA duplexes containing different single mismatched base pairs. Nucleic Acids Res 15:797–811

Jeffreys AJ, Wilson V, Thein SL (1985) Individual-specific 'fingerprints' of human DNA. Nature 316:76–79

Lander ES (1989) DNA fingerprinting on trial. Nature 339: 501–505

Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, Martin C, et al (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping. Science 235:1616–1622

Orrego C, King MC (1990) Determination of familial relationships. In: Innis MA, Gelfand DH, Sninsky JJ, White TJ (eds) PCR protocols. Academic Press, San Diego, pp 416–426

Robin ED, Wong R (1988) Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells. J Cell Physiol 136:507–513

Saiki RK, Bugawan TL, Horn GT, Mullis KB, Erlich HA (1986) Analysis of enzymatically amplified β-globin and HLA-DQα DNA with allele-specific oligonucleotide probes. Nature 324:163–166

Saiki RK, Walsh PS, Levenson CH, Erlich HA (1989) Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes. Proc Natl Acad Sci USA 86:6230–6234

Sokal RR, Rohlf FJ (1981) Biometry. WH Freeman, New York

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Vigilant L, Pennington R, Harpending H, Kocher TD, Wilson AC (1989) Mitochondrial DNA sequences in single hairs from a southern African population. Proc Natl Acad Sci USA 86:9350–9354

von Beroldingen CH, Blake ET, Higuchi R, Sensabaugh GF, Erlich HA (1989) Applications of PCR to the analysis of biological evidence. In: Erlich HA (ed) PCR technology. Stockton, New York, pp 209–223

Walberg MW, Clayton DA (1981) Sequence and properties of the human KB cell and mouse L cell D-loop regions of mitochondrial DNA. Nucleic Acids Res 9:5411–5421

Watterson GA (1975) On the number of segregating sites in genetic models without recombination. Theor Popul Biol 7:256–276

Whittam TS, Clark AG, Stoneking M, Cann RL, Wilson AC (1986) Allelic variation in human mitochondrial genes based on patterns of restriction site polymorphism. Proc Natl Acad Sci USA 83:9611–9615