ELSEVIER

Gene 205 (1997) 161–171

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

# Nuclear DNA diversity in worldwide distributed human populations

Ewa Ziętkiewicz [a], Vania Yotova [a], Michal Jarnik [a,1], Maria Korab-Laskowska [a,2], Kenneth K. Kidd [b], David Modiano [c], Rosaria Scozzari [d], Mark Stoneking [e], Sarah Tishkoff [b,3], Mark Batzer [f], Damian Labuda [a,*]

[a] *Centre de Recherche de l'Hôpital Sainte-Justine, Centre de Cancérologie Charles Bruneau, Département de Pédiatrie de l'Université de Montréal, Montréal, (Québec), Canada H3T 1C5*

[b] *Yale University School of Medicine, Department of Genetics, 333 Cedar St., New Haven, CT 06510, USA*

[c] *Fondazione Pasteur Cenci-Bolognetti, Istituto di Parassitologia, Universita 'La Sapienza', P. le A. Moro 5, 00185 Rome, Italy*

[d] *Dip. Genetica e Biologia Molecolare, Universita 'La Sapienza', P. le A. Moro 5, 00185, Rome, Italy*

[e] *Department of Anthropology, Pennsylvania State University, University Park, PA 16802, USA*

[f] *Department of Pathology, Stanley S. Scott Cancer Center, Louisiana State University Medical Center, 1901 Perdido Street, New Orleans, LA 70112, USA*

Accepted 14 July 1997

## Abstract

Nucleotide variation was examined in an 8 kb intronic DNA bordering exon 44 of the human *dystrophin* gene on Xp21. Thirty-six polymorphisms (substitutions, small insertions/deletions and one $(T)_n$ microsatellite) were found using SSCP/heteroduplex analysis of DNA samples from mixed Europeans, Papua New Guineans as well as from six African, three Asian and two Amerindian populations. In this way the European bias in the nuclear polymorphism ascertainment has been avoided. In a maximum likelihood tree constructed from the frequency data, Africans clustered separately from the non-African populations. Fifteen polymorphisms were shared among most of the populations compared, whereas 13 sites were found to be endemic to Africans and four to non-Africans. The common sites contributed most to the average heterozygosity ($H_n = 0.101\% \pm 0.023$), whereas the endemic ones, being rare, had little effect on this estimate. The $F_{ST}$ values were lower for Africans (0.072) than for non-Africans (0.158), suggesting a higher level of gene exchange within Africa, corroborating the observation of a greater number of segregating sites on this continent than elsewhere. The data suggest a recent common origin of the African and non-African populations, where a greater geographical isolation of the latter resulted in a smaller number of newly acquired polymorphisms.

*Keywords:* DNA polymorphism; Dystrophin; Human evolution; Ascertainment bias

## 1. Introduction

Disclosure of the content and organization of the genetic information in human DNA is a goal of the ongoing genome project. The linear sequence of nucleotides of the DNA message will soon be completely known. The second dimension of the human genetic heritage, its sequence variability within populations, has only started to be explored, both in qualitative and quantitative terms. According to the neutral theory (Kimura, 1983) most of the intra-specific variability at the molecular level is selectively neutral. It is maintained within the species by the balance between the mutational input and random extinction. The DNA polymorphisms represent a transient situation where one allele chosen at random is on its way to fixation. In the absence of selection the amount of genetic variation among subpopulations is shaped by mutation rate, random fluctuations in allele frequences (drift) and gene flow (exchange of alleles between populations), and strongly depends on demography (effective population size, migration

* Corresponding author. Tel: +1 514 345 4931, ext. 3586/sec. 3282; Fax: +1 514 345 4731; e-mail: Labuda@ere.Umontreal.ca

[1] Present address: NIAMS, LSBR, Bethesda, USA

[2] Present address: Département de Biochimie, Université de Montréal, Montréal, Canada

[3] Present address: Department of Biology, Pennsylvania State University, University Park, PA 16802, USA

Abbreviations: PCR, polymerase chain reaction; RFLP, restriction fragment length polymorphism; SSCP, single-strand conformational polymorphism.

*PII* S0378-1119(97)00408-3

etc.). For this reason the genomic variability observed in the present-day human population is not only a record of the history of our genes but also of our species; knowing it better is crucial to solving the questions concerning the evolution of modern humans.

The knowledge of human DNA diversity originates to a large extent from studies using markers developed for linkage mapping and ascertained in a small number of samples, primarily of European origin. As a result of the European ascertainment, the human genomic variability has suffered from the frequency bias of the polymorphisms tested (e.g. Bowcock et al., 1991; Mountain and Cavalli-Sforza, 1994; Rogers and Jorde, 1996). Furthermore, the use of RFLP markers was a source of sequence-based bias, since recognition sites of frequently used restriction enzymes often included CpGs of which the human genome is depleted (Bird, 1987; Labuda and Striker, 1989), and had an average G+C content that did not match the base composition of human DNA (see e.g. Cooper and Schmidtke, 1984; Cooper et al., 1985).

As a result of the limited sampling of individuals and/or populations as well as the differences in detection methods, the estimates of genome variability differed significantly. The density of segregating sites on the autosomes ranged from 1/370 to 1/108 (e.g. Murray et al., 1984; Cooper et al., 1985; Chakravarti et al., 1986; Hofker et al., 1986; Antonarakis et al., 1988; Fullerton et al., 1994), whereas the values reported for the X chromosome were from 1/732 to 1/190 (Cooper et al., 1985; Hofker et al., 1986; Zietkiewicz et al., 1995). The estimates of nucleotide diversity $H_n$ ranged from 0.1 to 0.39% on the autosomes (Murray et al., 1984; Cooper et al., 1985; Chakravarti et al., 1986; Hofker et al., 1986; Antonarakis et al., 1988; Li and Sadler, 1991; Fullerton et al., 1994) and from 0.04 to 0.09% on the X (Cooper et al., 1985; Hofker et al., 1986; Zietkiewicz et al., 1995).

To obtain unbiased information on the human genome diversity we investigated sequence variability in 13 globally distributed human populations. An 8 kb DNA segment of the human X chromosome was examined using SSCP combined with the heteroduplex analysis, a direct and objective technique comparable to DNA sequencing but more cost-effective (Orita et al., 1989; Zietkiewicz et al., 1992). The analysis of the distribution of the new (non-ancestral) allele frequencies at the 35 segregating sites in all the populations studied provided us with new information, shedding light on the history of modern humans.

## 2. Materials and methods

### 2.1. Genomic samples

Human DNA samples from the existing collections in the authors' laboratories are listed in Table 1.

### 2.2. Dys44 *sequence polymorphisms*

We analysed a total of 7622 nucleotides out of the 8035 nucleotide-long *dys44* sequence, spanning exon 44 of the dystrophin gene (cDNA positions 6499–6646) and its flanking introns between positions −2853 to −1 upstream and 1 to 5034 downstream. The numbering refers to the GenBank sequence U94396, where the following modifications to the previously deposited sequence (M81257) were introduced: a G-to-A change at position 710 of intron 44, a one-nucleotide insertion at position 1754 of intron 44, and a 195-nucleotide 'insertion' including positions −2347 to −2541 (replacing a 10-nucleotide segment −2347 to −2356) in intron 43. To search for gel mobility variants the *dys44* segment was divided into 30 fragments ranging from 92 to 480 bp, obtained by PCR combined where indicated with restriction enzyme digestion (Fig. 1 and Table 2). The amplification was in 20 µl, in 20 mM Tris–HCl (pH 8.4) (at 25°C), 50 mM KCl, 1.5 mM $MgCl_2$, 0.2 mM dNTPs (each), 0.2 µl (2 µCi) of α-[$^{32}$P]dCTP, with 1 µM primers (each), 1 unit Taq DNA polymerase (BRL) and 5 ng DNA. Following denaturation (3 min, 94°C), the amplification was performed for 30–35 cycles (94°C for 30 s, annealing temperture as in Table 2 for 45 s and 72°C for 90 s), terminated at 72°C (10 min). The PCR products were examined using SSCP/heteroduplex analysis, as described (Zietkiewicz et al., 1992). For typing length variants the PCR step at 72°C was 30 s rather then 90 s, the primer was radioactively labelled at its 5′-end (see Zietkiewicz et al., 1992 and below), and the electrophoresis was carried out under denaturing conditions. DNA mobility variants were characterized by direct dideoxy-sequencing of the PCR products (ABI 373A DNA sequencer).

### 2.3. *Allele specific oligonucleotide, ASO, hybridization*

Pentadecanucleotides were synthesized complementary to each allelic variant identified, to serve as ASO-probes for DNA typing by dot-blot hybridization (Table 3). 2–8 kb (kilobase) *dys44* fragments (depending on DNA sample quality) were amplified as above using combinations of primers from Table 2, except that the volume was 50 µl, the radioactivity was omitted, and a two-step PCR-program was applied (first 14 cycles were as above but the 72°C incubation was 120 s; in subsequent 25 cycles the 72°C incubation time increased in 7 s increments per cycle). Each reaction was brought up to 100 µl with water, incubated at 94°C for 2 min and cooled on ice; the solution was then made 10 × SSC (1.5 M NaCl, 150 mM sodium citrate, pH 7) by addition of 100 µl 20 × SSC. Aliquots of 100 µl were then applied in parallel onto a Hybond㋲-N$^+$ membrane (Amersham) to create two identical twin 48-dot blots. After rinsing with 100 µl of 10 × SSC the membranes were

Table 1
Population parameters for the 13 groups studied

| Population | Number of chromosomes | $S$ | $H_n \times 10^3$ (±S.E.) | $H_{nEuro} \times 10^3$ | $H_m$ |
|---|---|---|---|---|---|
| African Americans | 67 | 31 | 1.06 (±0.23) | 0.91 | 0.813 |
| Mossi (Burkina Faso) | 25 | 24 | 1.05 (±0.23) | 0.81 | 0.883 |
| Rimaibe (Burkina Faso) | 23 | 24 | 1.06 (±0.23) | 0.79 | 0.786 |
| Biaka Pygmies (Central Africa Republic) | 89 | 21 | 0.98 (±0.23) | 0.89 | 0.618 |
| Nigerians | 15 | 21 | 0.92 (±0.22) | 0.76 | 0.558 |
| M'Buti Pygmies (Congo) | 58 | 18 | 0.82 (±0.21) | 0.63 | 0.517 |
| Europeans (mixed) | 175 | 20 | 0.74 (±0.18) | 0.74 | 0.394 |
| Siberians[a] | 24 | 19 | 0.93 (±0.22) | 0.93 | 0.360 |
| Papua New Guineans[b] | 69 | 16 | 0.71 (±0.20) | 0.70 | 0.205 |
| Chinese | 85 | 18 | 0.76 (±0.20) | 0.76 | 0.131 |
| Japanese | 67 | 16 | 0.70 (±0.20) | 0.69 | 0.061 |
| Maya | 80 | 16 | 0.84 (±0.22) | 0.84 | 0.353 |
| Karitiana (Brazil) | 83 | 15 | 0.60 (±0.16) | 0.60 | 0.281 |
| World | 860 | 35 | 1.01 (±0.23) | 0.90 | 0.579 |
| Africans | 277 | 33 | 1.05 (±0.23) | 0.84 | 0.773 |
| Non-Africans | 583 | 22 | 0.90 (±0.23) | 0.90 | 0.270 |

$S$, Number of segregating sites (including substitutions and small indels); $H_n$, *average heterozygosity per n*ucleotide site (nucleotide diversity); $H_{nEur}$, '$H_n$' taking into account only the polymorphisms found in Europeans; $H_m$, *average heterozygosity of the* $(T)_n$ *m*icrosatelite; [a]Tundra and Forest Nentsi; [b]Coastal and Highland.

immersed in 1.5 M NaCl, 0.5 M NaOH for 10 min and in 1.5 M NaCl, 0.5 M Tris–HCl (pH 7.2) for 15 min DNA was fixed by exposing the membranes to 254 nm UV at the energy of 120 000 μJ/cm$^2$ in Stratalinker 1800 (Stratagene).

Blots were pre-hybridized for 30 min (rotary oven) in 20 ml 1 × SSPE (150 mM NaCl, 10 mM $NaH_2PO$, 1.1 mM EDTA, pH 7.4), 0.75 M NaCl, 70 mM Tris–HCl (pH 7.4), containing 1% SDS and 200 μg/ml heparin, at hybridization temperature, $T_H$ (see Table 3). ASO probes, 50 pmol, were 5′-labelled using $\gamma$-[$^{32}$P]ATP (6000 Ci/mmol) and T4 kinase (Gibco BRL) to a specific activity of 1–3 × 10$^6$ cpm/pmol (250 000–750 000 cpm/ng). Hybridization with the 0.8–2.0 pmol ASO probe (~2 000 000 cpm) was carried out for 40 min at $T_H$. The membranes were then washed with 2 × SSPE, 0.1% SDS for 10 min at room temperature, two times for 10 min at $T_H$, rinsed with 2 × SSC at room temperature, and exposed overnight at −80°C with a screen. Identical twin membranes were probed by the allelic ASOs and always read in parallel. This compensated, if necessary, for varying concentrations of the individual amplified DNA samples and, by the same token, for the variance in the probe activity. In addition, DNA samples of known allelic content (SSCP variants of known sequence) served as positive controls for the allelic probes. After stripping (5 min in boiling 0.5% SDS) the membranes were stored at −20°C and reused (up to 12 times) for hybridization with other probes.

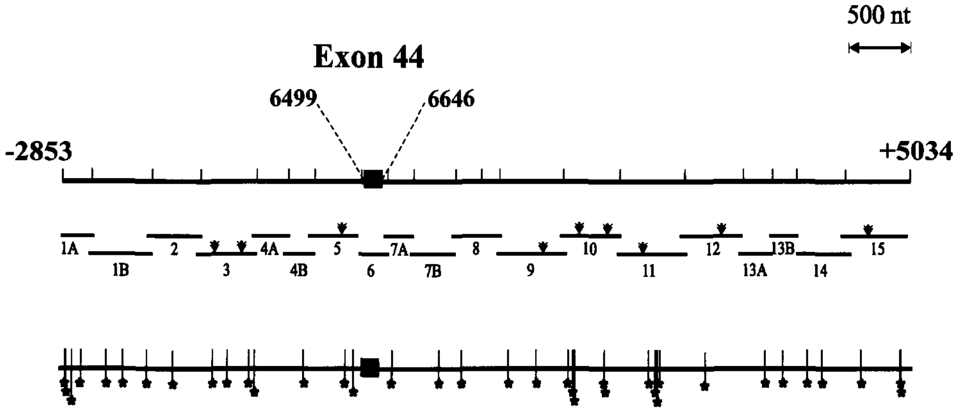

Fig. 1. *Dys44* DNA segment. Flanking nucleotides −2853 and 5034 delimit *dys44* region within the upstream and the downstream intron of dystrophin exon 44 (cDNA positions 6499–6646), respectively. Short horizontal lines below correspond to the overlapping PCR fragments (Table 2) used in the SSCP analysis, some of which were restriction digested (double-head arrows). The approximate distribution of the polymorphic sites (asterisks) is shown at the bottom.

Table 2
PCR amplification of *dys44* fragments

| *Dys44* fragment | PCR primers | | $T_A$ | RE | Fragment size (bp) |
|---|---|---|---|---|---|
| 1A | F | AGG GGG ATT TGT TGA AT | 52 | - | 296 |
| | R | GCA CAG AGA AGT ACC AGT T | | | |
| 1B | F | CAA GAG TGG AGA CTG ATG G | 58 | - | 427 |
| | R | GAC AGC ATT TTG GTA GCA T | | | |
| 2 | F | CTC TTC GGC TAC CTT CG | 52 | HindIII | 206, 294 |
| | R | TGG GCA GAG AAA GGA ATT A | | | |
| 3 | F | GTG CCT GGC TAT TAG TAA | 60 | HinfI | 197, 303, 92 |
| | R | CAT TTG GTC ACC TTC CAG T | | | |
| 4A | F | AAC ATA CAG CCC TGG TcC | 56 | - | 390 |
| | R | GTC ACA CTG TAC CCC ATA | | | |
| 4B | F | ACG AAT TAT TGA TTT ATT GG | 52 | - | 301 |
| | R | GTT CAT CAA CTG AAA GGA GT | | | |
| 5 | F | TGG TAA CTT TGT TCA TAT TA | 52 | AluI | 326, 145 |
| | R | GCG TAT ATT TTT TGG TTA TA | | | |
| 6 | F | CTT GAT CCA TAT GCT TTT ACC TGC A | 56 | - | 267 |
| | R | TCC ATC ACC CTT CAG AAC CTG ATC T | | | |
| 7A | F | TGA TTT GTT TTT TCG AAA T | 52 | - | 311 |
| | R | ACC TTG CTG TTA CGA TGC T | | | |
| 7B | F | AGC ATC GTA ACA GCA AGG | 56 | - | 444 |
| | R | CCG TGT AAT AAA CAC AGT G | | | |
| 8 | F | TTG CTA AAT TAC ATA GTT TAG GC | 52 | - | 329 |
| | R | AGC CCA AAA TTC ACT TC | | | |
| 9 | F | GTA TGG ATT CCC AAT CTG | 58 | EcoRI | 463, 169 |
| | R | TTA CAG AGA AAA GAG AGA CC | | | |
| 10 | F | TAG GCT CCT TAA AGT GCC | 56 | DdeI | 178, 251, 150 |
| | R | CTT ACA CTG TTC TGG TCA TAA C | | | |
| 11 | F | TTT TAT GCT TTG TTA TGA CC | 52 | HindIII | 265, 385 |
| 11A | F | GCA TAC ATG AAC TGA TCA AG | 52 | AluI | 44, 98, 285 |
| | R | ATG AGG ACC ATT AGA CAT TC | | | |
| 11POL | F | GTT GAA AAC CGT TGT AG | 48 | - | 85 |
| | R | GCT GAA CTT AAT CTC CT | | | |
| 12 | F | GAG AAT GTC TAA TGG TCC TC | 56 | HaeIII | 377, 255 |
| long12-15 | F | TCA GGC AGG AGA TTA AGT TCA | 54 | - | 2334 |
| | R | GGT ACA TTA TGT GCG TGT AT | | | |
| 13A | F | ATA CAC GCA CAT AAT GTA CC | 59 | - | 291 |
| | R | TAT ATG GGC TGC TTG GTT GT | | | |
| 13B | F | AGA CAG GAT GTA GAC AGT GG | 59 | - | 350 |
| | R | AGT TGC ATG GAA CCA GAT | | | |
| 14 | F | TCA TTC AAA ATA TAC TGA CTG | 52 | - | 490 |
| | R | GAG AAG AGA AAT CAG AGA CAT | | | |
| 15 | F | AAA TCC CAG GTC CTT TAG C | 56 | PstI | 262, 426 |
| | R | GTT ATT GTT TCT ACT GGC AAC T | | | |
| 8kb | F | AGT CTG TGC CAC AGG TTT GAA ATC GAA | 50 | - | 8181 |
| | F | CAG GCT TGT ATG TCT GCG ACA ACG TTC | | | |

$T_A$, Annealing temperature in °C; RE, restriction enzyme applied; F, forward; R, reverse.

### 2.4. Data analysis

Allele frequencies were determined by direct gene counting. A maximum likelihood tree was constructed using the CONTML program from Phylip package 3.5 (Felsenstein, 1993). Heterozygosity $h_i$ at each sequence site $i$ was calculated as: $h_i = \sum_{j=1}^{a} p_j(1-p_j)$, where $p_j$ is a frequency of the $j$th allele and $a$ is a number of alleles at the $i$th site. Microsatellite heterozygosity, $H_m$ for the $(\mathrm{T})_n$ site, was calculated accordingly and reported separately from $H_n$ calculated below. Average heterozygosity per nucleotide $H_n$ (nucleotide diversity; including here substitutions and small insertions/deletions) was calculated as a sum of heterozygosities $h_i$ divided by the total DNA length $L(L=7622)$: $H_n = \sum_{i=1}^{L} h_i/L$; standard error: $\mathrm{SE} = \left[\sum_{i=1}^{L} (h_i - H_n)^2/(L-1)L\right]^{1/2}$. The $H_n$ values for the groups of populations, such as Africans, non-Africans or the world, were obtained as above, using allele frequencies $p_j$ averaged across the constituting subpopulations. $F_{ST}$ statistics were calculated according to Hartl and Clark (1989): $F_{ST} = \dfrac{H_T - H_S}{H_T}$, where $H_T$ is the nucleotide diversity in the group of subpopulations and $H_S$ is computed by dividing the sum of nucleotide diversities of the constituting subpopulations by their number. $F_{ST}$ values for individual polymorphic sites were obtained using $h_i$ rather than $H_n$ values to calculate $H_T$ and $H_S$.

Table 3
Oligonucleotide probes used in ASO hybridization

| Polymorphic site | ASO sequence | $T_H$ | Polymorphic site | ASO sequence | $T_H$ |
|---|---|---|---|---|---|
| 2A-3A | F TCT GTG **AAA** CAG GTT | 37 | 45 | R GGC **c/a**TT AAA AAT TGG | 37 |
| 2C-3C | F TCT GTG **CCA** CAG GTT | | | | |
| 2C-3A-5 | F TCT GTG **CAA** CA**g/a** GTT | | 48 | R TCT ATA TCT A**c/t**C CCA | 37 |
| 8 | F AGT AGC TAA **a/c**AG TGT | 37 | 50 | F AGG TTA **Ac/t**T AGG GAG | 37 |
| 10 | F TTC **t/c**GA CTC TCA ATA | 37 | 55 | F CCA TGT AGT At**a/g** TAT | 37 |
| 12 | F Tg**a/g** TGC AAA TGC ATG | 38 | 58T-59 | R CCA A**c/a**T T**T**T ACT TCT | 37 |
| | | | 58A-59 | R CCA A**c/a**T T**A**T ACT TCT | |
| 14 | F CTT **c/t**GG CTA CCT TCG | 38 | | | |
| | | | 64A-65 | F AAT ATT CTC A**Ac/a** CCT | 37 |
| 15 | F GAA ACA AGA GA**t/g** ATT | 38 | 64C-65 | R AGG **t/gC**T GAG AAT ATT | |
| 18 | F ATC CA**t/c** TGC TTC TTA | 38 | 70 | F TTT AAC A**g/a**A AGC CTC | 37 |
| 20 | F ATG CTT GCT TGG **a/g**CA | 43 | 71 | F AAC CG**t/c** TGT AGC ATA | 37 |
| 25 | F ATG C**c/g**C CAG TTG ATT | 38 | 72 | R AGA CAA **a/t**AT ATA TGC | 37 |
| 30 | F AGG TAA ACA TA**c/t** AGC | 43 | 85 | F CTG GTC TAG ATC TC- --- --- -C | 37 |
| | | | | F G ATC TC**t aga tct c**C | |
| 32 | F ATA A**t/c**T TCT CTG TGG | 37 | | | |
| | | | 86 | F GTG AAA AAA **a/g**TC AAC | 37 |
| 33 | F AAT GT**t gt**G TGT ACA | 37 | | | |
| | F AAT GT- --G TGT ACA TGC | | 87 | F TGG AGA A**c/t**C TCT CTA | 37 |
| 35 | R AGT TTC **c/t**TG CAT TTG | 37 | 88 | F CAT TTT TTT ACC **t/c**GA | 37 |
| 38 | F GTT GTC **atc** ATT ATA | 37 | 90 | F ATC CAC CAA AA**c/t** AGT | 38 |
| | F GTT GTC --- ATT ATA TTA | | | | |
| | | | 93 | F ACT GAA TTA T**c/t**G TCT | 38 |
| 40 | F ATA G**a/g**G AAA CAG CAT | 37 | | | |
| | | | 95 | F CTA T**g/a/c**A TAG AAT GAG | 35 |

$T_H$, Hybridization temperature in °C; F and R, oligonucleotides complementary to the − and + strand, respectively; allelic positions within the oligonucleotide sequence are in bold; whenever possible allelic probes are shown as a single sequence within variants in lower case separated by '/'.

## 3. Results and discussion

### 3.1. DNA sequence variants and their frequency profiles in populations

An 8 kb region spanning exon 44 in the dystrophin gene (*dys44*, Fig. 1) was examined for the presence of DNA polymorphisms by combined SSCP/heteroduplex analysis (Fig. 2(a)). On average 20 chromosomes from each of the 13 subpopulations listed in Table 1 were investigated, which amounts to a survey of almost 2 Mb DNA for the presence of mutations. ASO hybridization (Fig. 2(b)) was used to determine allele frequencies in the extended, global population sample (Table 1). The characterization of gel mobility variants (Fig. 2(a)) by DNA sequencing revealed a nine-allele $(T)_{14-23}$ length variation (Fig. 2(c)) and 35 'point mutation' polymorphisms due to 11 transversions, 22 transitions, two

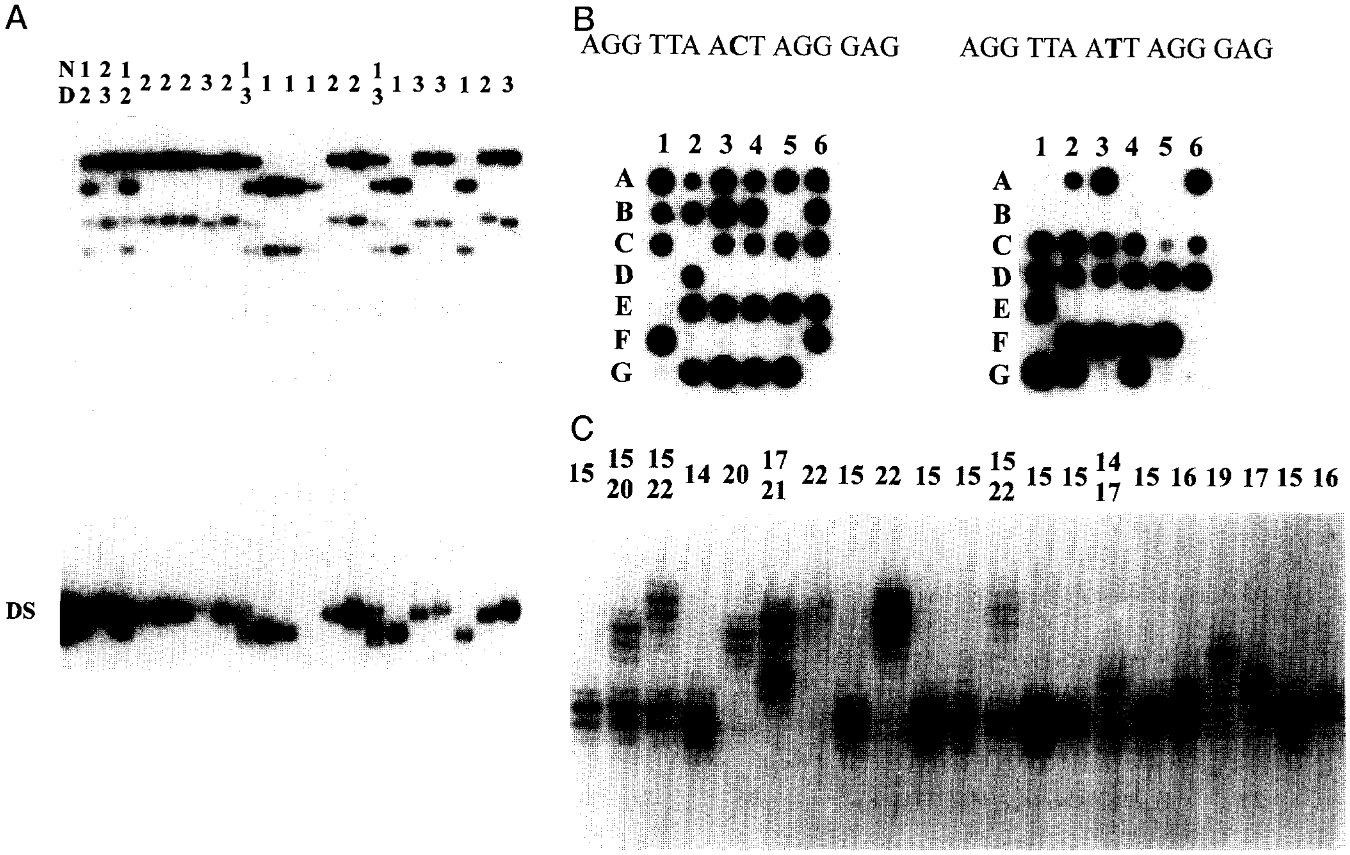

Fig. 2. Polymorphism detection and genotyping. (A) SSCP/heteroduplex analysis (moblity variants are denoted at the top, ND is the non-denatured control, DS indicates migration of double-stranded DNA); (B) ASO dot-blot hybridization (positions B5 and G6 correspond to 'no DNA' controls; the sequences of ASO probes are shown at the top); (C) denaturing gel electrophoresis of $(T)_n$ variants ($n$ is indicated at the top).

3-nucleotide deletions and one 8-nucleotide duplication (Table 3). The ratio of substitutions to small insertions/deletions was approximately 10, and that of transitions to transversions was 2.

Frequency data (excluding length polymorphism) were used to construct a maximum likelihood tree (Felsenstein, 1993): the African populations clustered together separately from the non-African ones (Fig. 3). The ancestral state of each polymorphic site was inferred by comparison with the orthologous non-human primates: allele shared with at least two great apes, i.e. chimpanzee, gorilla and/or orangutan, was considered ancestral (Zietkiewicz et al., submitted). A theoretical ancestral population obtained by setting ancestral allele frequencies at 1.0 (see Batzer et al., 1994; Mountain and Cavalli-Sforza, 1994; Nei and Takezaki, 1996) joined the tree between M'Buti Pygmies and other African populations, consistent with the African origin of modern humans (Stringer and Andrews, 1988; Lahr and Foley, 1994; Tattersal, 1995).

The number of segregating sites (Table 1) was typically higher in African subpopulations than in non-African ones and ranged from 15 in Karitiana, a small inbred Amerindian tribe (Kidd et al., 1991), to 31 in African Americans, a highly heterogeneous and admixed group (Chakraborty et al., 1992). The frequency profiles of polymorphisms in populations are compared in a

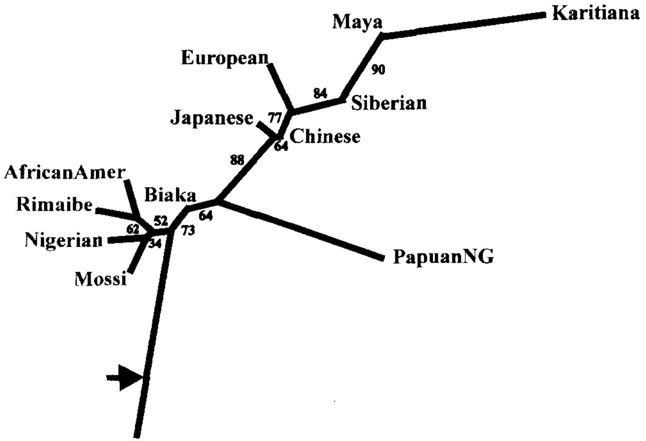

Fig. 3. Maximum likelihood tree of human populations. The genetic distances are proportional to the branch lengths; bootstrap values (100 replicates) are shown at the branching points; the putative root (indicated by an arrow) was obtained by setting the ancestral allele frequencies at one (see text).

series of histograms in Fig. 4. Fifteen segregating sites, with an overall new (i.e. non-ancestral) allele frequency ≥0.1 were observed in all the populations. We conclude that these polymorphisms had to be present in an ancestral population, which gave rise to both African and non-African groups. Thirteen polymorphisms were endemic to Africa and only four such sites were found outside Africa (new alleles at positions '10' and '85', frequent in Europeans, Siberians and Amerindians, always occurred together; in African Americans they were considered as European admixture). These African and non-African polymorphisms were rare, with new allele frequencies ≤0.1. Four out of the 13 African and two out of the four non-African polymorphisms were 'private', i.e. a new allele was present in a single population only (indicated by asterisks in Fig. 4). The polymorphism at site '55' was found only in Europeans and African Americans. A new allele was fixed at site '30'

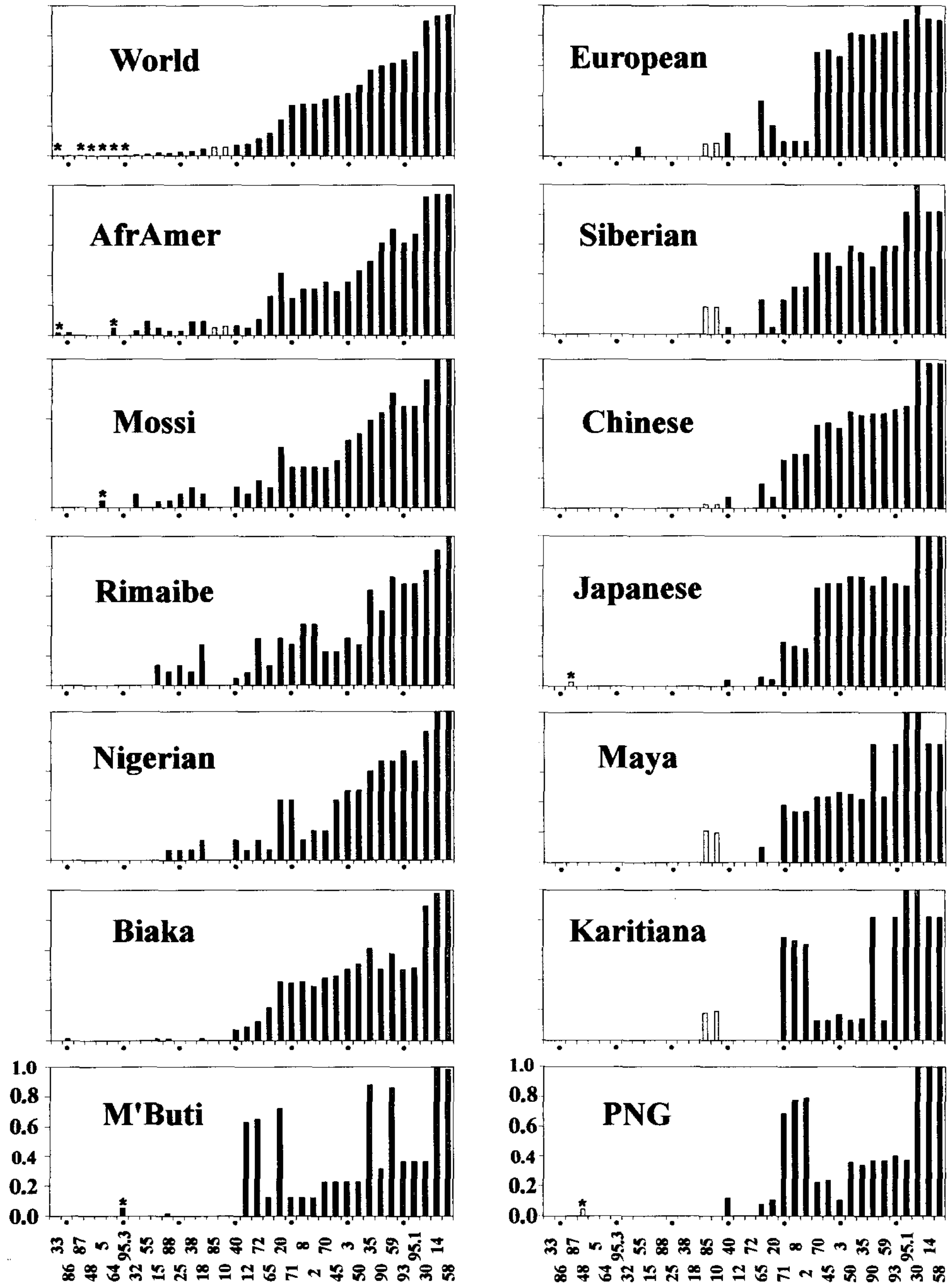

Fig. 4. Distribution of *dys44* polymorphisms among populations. Histograms of new allele frequencies are ordered according to their increasing abundance in the world population. Globally distributed polymorphisms are represented by dark grey bars, endemic African *positions in black and those found only in non-Africans in light grey. Asterisks indicate polymorphisms found in a single population. Arbitrary names of the polymorphisms (Table 3) are shown at the bottom histograms only.

in all non-African populations; at site '14' this occurred in some of the African populations, whereas at site '58' in a number of African and non-African populations (Fig. 4).

In Africans the profiles of new allele frequencies (Fig. 4) are similar in all the populations, except for the most distant M'Buti Pygmies (Fig. 3). In Biaka Pygmies we noted the shortage of rare new alleles; in Mossi, Rimaibe, Nigerians and Biaka Pygmies the new allele was fixed at position '58' whereas in Mossi, Nigerians and M'Buti Pygmies this occurred at position '14'. In non-Africans, Fig. 4 reveals differences among populations at a number of positions. In Chinese, Japanese and Papua New Guineans the new alleles '95–1' and '14' are fixed whereas those at positions '10' and '85' are missing. Papua New Guineans share with the Amerindian populations a higher level of new alleles at positions '8', '2' and '71'. Maya and Karitiana also share new alleles at positions '10' and '85'; these are also present at significant frequency in Europeans and Siberians. Otherwise the frequency profiles in Amerindians appear different; the distinct one in Karitiana may be ascribed to the effect of the genetic drift expected to be more pronounced in a small and inbred population (Kidd et al., 1991). Due to their physical proximity *dys44* polymorphisms are in linkage disequilibrium, which can be seen as correlated changes in certain allele frequencies within populations. However, based on tests by Tajima (1989) and Kimura and Crow (1964), these sites appeared as independent, neutral polymorphisms (data not shown).

Four to nine alleles were observed at the $(T)_n$ microsatellite in African populations and from two to four in non-Africans (Fig. 5); a similar pattern to that seen in Fig. 4, but even more pronounced. The $(T)_{15}$ allele had the highest frequency in all but one (M'Buti Pygmies) population; it could be considered the oldest or the most stable among the inter-converting length variants.

### 3.2. $F_{ST}$

$F_{ST}$ values, reflecting the contribution of differences among subpopulations to the total divergence, strongly depend on how these subpopulations are defined within the total population (Fig. 6). The $F_{ST}$ value of 0.147 characterized the world composed of 13 subpopulations. When the world population was considered to be composed of only two groups, Africans and non-Africans, the corresponding value was 0.036. This indicated that most of the divergence within the world was caused by differences among particular populations and not by the difference between Africans and non-Africans. Furthermore, the $F_{ST}$ values for Africans and non-Africans were 0.072 and 0.158, respectively. The same overall tendencies were observed when $F_{ST}$ was calculated for each polymorphic site separately (Fig. 6). In non-Africans the $F_{ST}$ values were much higher at the sites considered to represent old polymorphisms, whereas in Africans the $F_{ST}$ values were similar across most of the polymorphic sites, except for the African-specific positions '12', '72' and '30'.

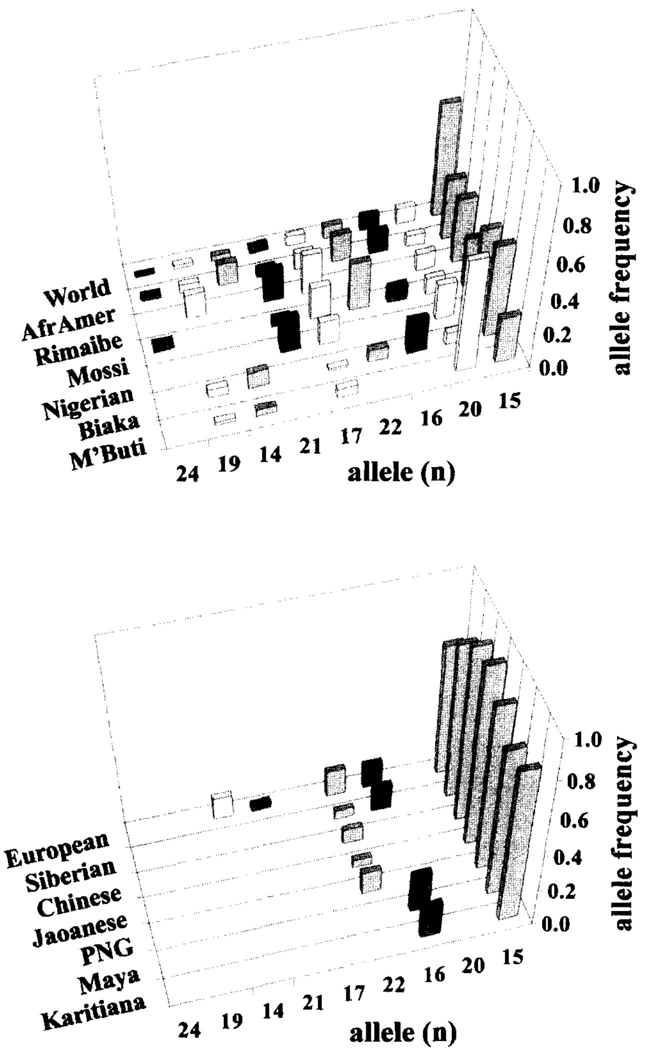

Fig. 5. Allele frequencies of the $(T)_n$ microsatellite among Africans and non-African populations ($n$, 14 to 24, are shown at the bottom).

In Africans the major contribution to their overall divergence comes from M'Buti Pygmies: excluding this population resulted in lowering the African $F_{ST}$ from 0.072 to 0.027. Among non-Africans the two groups contributing most to the $F_{ST}$ value were Papua New Guineans and Karitiana. When both these populations were omitted from the calculation, the $F_{ST}$ dropped from 0.158 to 0.059 (excluding only Papua New Guineans or Karitiana resulted in $F_{ST}$ of 0.127 or 0.121, respectively). Nevertheless, in spite of these different representations of non-Africans and Africans, $F_{ST}$ remained higher outside Africa. This is understandable in the light of the difference in geographical areas occupied by Africans and non-Africans.

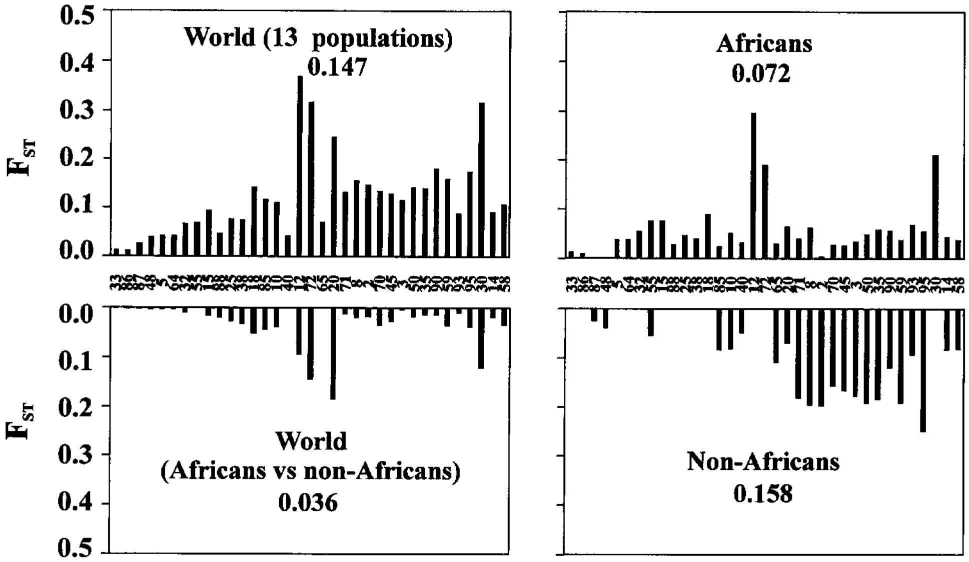

Fig. 6. $F_{ST}$ values at each of the 35 polymorphic positions (bars) and based on the average heterozygosty per nucleotide (numbers). (a) World composed of the 13 populations studied; (b) world composed of two groups, Africans and non-Africans; (c) Africans (six populations); and (d) non-Africans (seven populations).

In conclusion, our $F_{ST}$ data are consistent with earlier reports on a variety of markers (Relethford, 1995; Hammer et al., 1997) and suggest a higher level of gene exchange within Africa than among non-African populations. This in turn may explain the greater number of polymorphic sites seen in Africa. However, the same effect could also result from a larger population size on the Old Continent (Relethford and Harpending, 1995; Stoneking et al., submitted). On the other hand, higher nuclear DNA variability among Africans cannot be taken by itself as an indication of the African origin of modern humans. African and non-African populations share all old segregating sites and differ in rare, new polymorphisms (as measured by the frequency of the new, non-ancestral allele). These results are therefore consistent with a scenario where Africans and non-Africans diverged recently from a common ancestral population, with more opportunities (see above) to accumulate new variants existing subsequently in Africa than outside (Zietkiewicz et al., submitted).

### 3.3. Nucleotide diversity and the ascertainment bias

The density of segregating sites reported previously for the X chromosome ranged from 1/732 to 1/190 (Cooper et al., 1985; Hofker et al., 1986; Zietkiewicz et al., 1995). In *dys44*, with 35 sites within 7622 bp, this density is 1/218 and is probably an underestimate given the limitations of SSCP/heteroduplex analysis (Hayashi, 1992; Sheffield et al., 1993). Interestingly, using 34 restriction enzymes commonly applied in the earlier RFLP studies (e.g. Cooper and Schmidtke, 1984; Cooper et al., 1985; Antonarakis et al., 1988), five variant positions would be revealed within 540 bp of *dys44* screened in that way. The resulting polymorphism density of 1/108 would therefore be twice as large as our SSCP-based estimate. If only Europeans were analysed, the RFLP density would have been 1/180, compared to 1/382 (20/7622) based on the SSCP screening of the same population.

In our data set the $H_n$ values follow the same relative tendency as the number of polymorphic sites $S$ (Table 1). $H_n$ is estimated at 0.101% for the world and its highest values are seen in Africans. When *dys44* polymorphisms are ascertained in European DNAs only (Fig. 4), the resulting biased heterozygosity values $H_{nEur}$ are lower. A significant effect of such a bias is only seen in African populations (Table 1); but even there, it does not exceed 25%. This is because European polymorphisms represent practically all common frequent sites, whereas rare new polymorphisms, particularly abundant in Africa, increase $S$ but contribute little to the overall heterozygosity. However, if only European markers are considered, the ratio of nucleotide diversity in Africa and non-Africa becomes inverted (in Africans $H_{nEur}$ of 0.084% is lower than in non-Africans, 0.09%). Furthermore, by typing only European markers we inadvertently lose the qualitative and quantitative information concerning polymorphic sites endemic to Africans or to other non-European populations (Fig. 4).

The knowledge of such polymorphisms is necessary to reconstruct the evolutionary history of our species. Those that can be considred 'private', may be important for detecting gene flow and reconstructing demographic history and relations among particular populations. Unbiased ascertainment of the polymorphisms thus appears to be essential to recognize the differences among the diversified human groups. In contrast to early nuclear data, mitochondrial DNA studies from the beginning included globally distributed human samples (Cann et al., 1987); this can be at least partly responsible for the conflicting portraits of human origins emerging from these two types of data (Hey, 1997). The microsatellites (Bowcock et al., 1994) and minisatellites (Armour et al., 1996) do not suffer from ascertaiment bias to the same extent (Rogers and Jorde, 1996), since these sites are often conserved across species (Zietkiewicz et al., 1994) so that their genomic distribution as polymoprhic sites is usually pan-specific. The microsatellite heterozygosity $H_m$ values in *dys44* (Table 1) are higher in Africa; this is related to the greater number of $(T)_n$ alleles, and resembles the nucleotide diversity in a sense that finding a new microsatellite allele is like adding a new polymorphic site in the analysed sequence segment. However, we need nuclear markers with different characteristics, and ideally their combinations (Batzer et al., 1994, Tishkoff et al., 1996) to study the evolutionary origin of *Homo sapiens*, which thus requires samples from different and globally distributed human populations (Cavalli-Sforza et al., 1991; Cavalli-Sforza et al., 1994).

## 4. Conclusions

(1) European bias leads to an inadvertent loss of the important genomic record concerning the recent evolution of our species. Therefore, it is essential to ascertain nuclear polymorphisms in a worldwide-distributed human population.

(2) Our data are consistent with the recent common origin of modern humans and indicate that differences between Africans and non-Africans are a result of the recent history of these populations. African populations staying on a single continent remained in closer contact, whereas those outside, being dispersed over larger geographical areas, were more isolated, which reduced opportunities for exchange of the genetic material.

## Acknowledgement

Wojciech Makalowski and Nur Akalin participated in primer design and early experiments, Guy Cardinal and Miguel Angel Roman in DNA sequencing. This research was supported by a grant from the Canadian Genome Analysis and Technology Program (GO-12782).

## References

Antonarakis, S.E., Oettgen, P., Chakravarti, A., Halloran, S.L., Hudson, R.R., Feisee, L., Karathanasis, S.K., 1988. DNA polymorphism haplotypes of the human apolipoprotein APOA1-APOC3-APOA4 gene cluster. Hum. Genet. 80, 265–273.

Armour, J.A.L., Anttinen, T., May, C.A., Vega, E.E., Sajantila, A., Kidd, J.R., Kidd, K.K., Bertranpeti, J., Paabo, S., Jeffreys, A.J., 1996. Minisatellite diversity supports a recent African origin for modern humans. Nature Genet. 13, 154–160.

Batzer, M.A., Stoneking, A., Alegria-Hartman, M., Bazan, H., Kass, D.Y., Shaikh, T.H., Novick, G.E., Ioannou, P.A., Scheer, W.D., Herrera, R.J., Deininger, P.L., 1994. African origin of human specific polymorphic *Alu* insertions. Proc. Natl. Acad. Sci. USA 91, 12288–12292.

Bird, P.A., 1987. GpG islands as gene markers in the vertebrate nucleus. Trends Genet. 3, 342–347.

Bowcock, A.M., Kidd, J., Mountain, J.L., Hebert, J.M., Carotenuto, L., Kidd, K.K., Cavalli-Sforza, L.L., 1991. Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. Proc. Natl. Acad. Sci. USA 88, 839–843.

Bowcock, A.M., Ruiz-Lineares, A., Tomfohrde, J., Minch, E., Kidd, J.R., Cavalli-Sforza, L.L., 1994. High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368, 455–457.

Cann, R.L., Stoneking, M., Wilson, A.C., 1987. Mitochondrial DNA and human evolution. Nature 325, 31–36.

Cavalli-Sforza, L.L., Wilson, A.C., Cantor, C.R., Cook-Deegan, R.M., King, M.-C., 1991. Call for a worldwide survey of human genetic diversity: a vanishing opportunity for the human genome project. Genomics 11, 490–491.

Cavalli-Sforza, L.L., Menozzi, P. and Piazza, A. (1994) *The history and geography of human genes.* Princeton University Press, Princeton, New Jersey.

Chakravarti, A., Elbein, S.C., Permutt, M.A., 1986. Evidence for increased recombination near the human insulin gene: implication for disease association studies. Proc. Natl. Acad. Sci. USA 83, 1045–1049.

Chakraborty, R., Kamboh, M.I., Nwankwo, M., Ferrell, R.E., 1992. Caucasian genes in American Blacks: new data. Am. J. Hum. Genet. 50, 145–155.

Cooper, D.N., Schmidtke, J., 1984. DNA restriction fragment length polymorphisms and heterozygosity in the human genome. Hum. Genet. 66, 1–16.

Cooper, D.N., Smith, B.A., Cooke, H.J., Niemann, S., Schmidtke, J., 1985. An estimate of unique DNA sequence heterozygosity in the human genome. Hum. Genet. 69, 201–205.

Felsenstein, J. (1993) PHYLIP (Phylogeny Inference Package) version 3.5p. Distributed by the author. Department of Genetics, University of Washington, Seattle. [Felsenstein, J. (1989) PHYLIP–Phylogeny Inference Package (version 3.2), *Cladistics*, **5**, 164–166.]

Fullerton, S.M., Harding, R.M., Boyce, A.J., Clegg, J.B., 1994. Molecular and population genetic analysis of allelic sequence diversity at the human $\beta$-globin locus. Proc. Natl. Acad. USA 91, 1805–1809.

Hammer, M.F., Spurdle, A.B., Karafet, T., Bonner, M.R., Wood, E.T., Novelletto, A., Malaspina, P., Mitchell, R.J., Horai, S., Jenkins, T., Zegura, S.L., 1997. The geographic distribution of human Y chromosome variation. Genetics 145, 787–805.

Hartl, D.L. and Clark, A.G. (1989) *Principles of Population Genetics*, Sinauer Associates, Sunderland, Massachusetts.

Hayashi, K., 1992. PCR-SSCP: A simple and sensitive method for detection of mutations in the genomic DNA. PCR Methods Appl. 1, 34–38.

Hey, J., 1997. Mitochondrial and nuclear genes present conflicting portraits of human origin. Mol. Biol. Evol. 14, 166–172.

Hofker, M.H., Skraastad, M.I., Bergen, A.A.B., Wapenaaar, M.C., Bakker, E., Millington-Ward, A., van Ommen, G.J.B., Pearson, P.L., 1986. The X chromosome shows less genetic variation at the restriction sites than the autosomes. Am. J. Hum. Genet. 39, 438–451.

Kidd, J.R., Black, F.L., Weiss, K.M., Balazs, I., Kidd, K.K., 1991. Studies of three Amerindian populations using nuclear DNA polyorphisms. Hum. Biol. 63, 775–794.

Kimura, M., 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

Kimura, M., Crow, J.F., 1964. The number of alleles that can be maintained in a finite population. Genetics 49, 725–738.

Labuda, D., Striker, G., 1989. Sequence conservation in Alu evolution. Nucleic Acids Res. 17, 2477–2491.

Lahr, M.M., Foley, R., 1994. Multiple dispersals and modern human origins. Evol. Anthropol. 3, 48–60.

Li, W.H., Sadler, L.A., 1991. Low nucleotide diversity in man. Genetics 129, 513–523.

Mountain, J.L., Cavalli-Sforza, L.L., 1994. Inference of human evolution through cladistic analysis of nuclear DNA restriction polymorphisms. Proc. Natl. Acad. Sci. USA 91, 6515–6519.

Murray, J.C., Mills, K.A., Demopulos, C.M., Hornung, S., Motulsky, A.G., 1984. Linkage disequilibrium and evolutionary relationships of DNA variants (restriction enzyme fragment length polymorphisms) at the serum albumin locus. Proc. Natl Acad. Sci. USA 81, 3486–3490.

Nei, M., Takezaki, N., 1996. The root of the phylogenetic tree of human populations. Mol. Biol. Evol. 13, 170–177.

Orita, M., Iwahana, H., Kanazawa, H., Hayashi, K., Sekiya, T., 1989. Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformational polymorphisms. Proc. Natl. Acad. Sci. USA 86, 2766–2770.

Relethford, J.H., 1995. Genetics and modern human origins. Evol. Anthrop. 4, 53–63.

Relethford, J.H., Harpending, H.C., 1995. Ancient differences in population size can mimic a recent African origin of modern humans. Curr. Anthropol. 36, 667–675.

Rogers, A., Jorde, L.B., 1996. Ascertainment bias in estimates of average heterozygosity. Am. J. Hum. Genet. 58, 1033–1041.

Sheffield, V.C., Beck, J.S., Kwitek, A.E., Sandstrom, D.W., Stone, E.M., 1993. The sensitivity of single-strand conformation polymorphism analysis for the detection of single base substitutions. Genomics 16, 325–332.

Stringer, C.B., Andrews, P., 1988. Genetic and fossil evidence for the origin of modern humans. Science 239, 1263–1268.

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

Tattersal, I. (1995) *The Fossil Trail.* Oxford University Press, New York, Oxford.

Tishkoff, S.A., Dietzsch, E., Speed, W., Pakstis, A.J., Kidd, J.R., Cheung, K., Bonné-Tamir, B., Santachiara-Benerecetti, A.S., Moral, P., Krings, M., Pääbo, S., Watson, E., Risch, N., Jenkins, T., Kidd, K.K., 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271, 1380–1387.

Zietkiewicz, E., Akalin, N., Labuda, D., 1995. Neutral polymorphisms in the deletion-prone regions of the dystrophin gene. Hum. Hered. 45, 80–88.

Zietkiewicz, E., Sinnett, D., Richer, C., Mitchell, G., Vanasse, M., Labuda, D., 1992. Single strand conformational polymoprhisms (SSCP): detection of useful polymorphisms at the dystrophin locus. Hum. Genet. 89, 453–456.

Zietkiewicz, E., Rafalski, A., Labuda, D., 1994. Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification. Genomics 20, 176–183.