Submitted to Journal of Molecular Evolution

Original Article

Minisatellite MS32 alleles show population specificity among Thai, Chinese and Japanese

Qing-Hua Yuan • Azusa Tanaka • Richard H. Kaszynski • Morio Iino • Tomoko Okuno • Tatsuaki Tsuruyama • Toshimichi Yamamoto • Alec J. Jeffreys • Keiji Tamaki

---

Q.-H. Yuan • A. Tanaka • R. H. Kaszynski • M. Iino • T. Okuno • T. Tsuruyama• K. Tamaki (✉)

Department of Forensic Medicine and Molecular Pathology, Graduate School of Medicine, Kyoto University Kyoto, 606-8501, Japan

T. Yamamoto

Department of Legal Medicine and Bioethics, Nagoya University Graduate School of Medicine, Nagoya, 466-8550, Japan

A. J. Jeffreys

Department of Genetics, University of Leicester, Leicester LE1 7RH, UK

Corresponding author: Keiji Tamaki

Tel; +81-75-753-4472; Fax +81-75-761-9591

e-mail:ktamaki@legal.med.kyoto-u.ac.jp

Running head: Population Specificity of Minisatellite MS32

Keywords: Minisatellite MS32 (D1S8), MVR-PCR, VNTR, mutation, human diversity

Abstract Lineages of structurally related alleles at minisatellite MS32 in human populations show considerable differentiation at the continental level. However, the regional specificity of these lineages remains unknown. We now describe the comparison of allele structures in Thai, Han Chinese and Japanese populations with lineages previously established for north Europeans and Africans. The great majority of alignable Asian alleles showed their closest structural relative in Asia, with few instances of preferential alignment of Asian with European alleles and only one isolated incident showing a best match with an African allele. Further, there was a strong tendency, most marked for Japanese, for Asian alleles to align preferentially with other alleles from the same population, indicating strong regional specificity of allele lineages. This rapidly evolving minisatellite can therefore serve as a lineage marker for exploring recent events in human population history and dissecting population structure at the fine-scale level, as well as being an extremely informative DNA marker for personal identification.

Key words Minisatellite MS32 (D1S8) • MVR-PCR • VNTR • mutation • human diversity

Introduction

Tandem repetitive DNAs, which comprise approximately 3% of the human genome, contain human minisatellites or variable number tandem repeat (VNTR) loci (International Human Genome Sequencing Consortium 2001). Human GC-rich minisatellites are preferentially found clustered in the recombination-proficient subtelomeric regions of chromosomes (Royle et al. 1988). The repeat unit length ranges from 6 to more than 100 bp, with arrays usually kilobases in length. Some minisatellite loci show very high levels of allele length variability. In addition, most loci consist of heterogeneous arrays containing two or more subtly different repeat unit types (minisatellite variant repeats, MVRs). For Example, the consensus 29-bp repeat unit

sequence of minisatellite MS32 (D1S8) (5'-G**R**CCAGGGGTGACTCAGAATGGAGCAGGY-3') shows two classes of repeat unit that differ by a single base substitution which creates or destroys a *Hae*III restriction site [designated **a**-type and **t**-type, respectively, (Jeffreys et al. 1990)]. These variant repeats can be mapped by the polymerase chain reaction (MVR-PCR) to provide a powerful approach for allele classification based on the interspersion patterns of variant repeats within the repeat array (Jeffreys et al. 1991). MVR-PCR followed by agarose gel electrophoresis and Southern blot hybridization allows such interspersion patterns to be displayed as ladders of PCR products. This approach has revealed enormous levels of allelic variation at several hypervariable minisatellites (Jeffreys et al. 1991; Neil and Jeffreys 1993; Armour et al. 1993; Buard and Vergnaud 1994; Andreassen and Olaisen 1998; Tamaki et al. 1999). At the D1S8 locus (minisatellite MS32), almost all alleles in several ethnic populations surveyed were different. However, different alleles can show significant similarities in repeat organization, implying recent common ancestry (Jeffreys et al. 1991). Heuristic dot-matrix algorithms have been developed to identify significantly related alleles and have shown that approximately three quarters of all alleles mapped to date in diverse populations can be grouped into over 100 sets of alignable alleles, indicating multiple relatively ancient lineages of related alleles present in diverse populations (Tamaki et al. 1995). Some groups of related alleles can display a strong tendency to be population-specific at the continental level, being restricted for example to Europeans, Asians or Africans and consistent with recent divergence from a common ancestral allele. However, the ability of these allele lineages to distinguish between populations at the sub-continental level has not been examined. We address this issue by expanding the current database of mapped MS32 alleles, including Japanese alleles, with additional alleles from Thais and Han Chinese and analyzing the lineages seen in these three Asian populations. While the advent of microsatellites has led to unprecedented progress in evolutionary research, it has in many ways marginalized minisatellites. The present study therefore exemplifies minisatellites as a highly effective tool for dissecting population structures at the very fine-scale level.

Materials and Methods

DNA samples

Thai (Bangkok) and Han Chinese (Beijing) DNA samples were kindly supplied by Prof. Y. Katsumata. DNA was extracted from the peripheral leukocytes of 170 healthy unrelated Thai volunteers and 93 Chinese volunteers with their informed consent. The concentrations of genomic DNA samples were quantified using a spectrophotometer (NanoDrop ND-1000, Scrum Inc., Tokyo). The present study, involving the use of the Thai and Chinese samples, was approved by the Ethics Committee of the Graduate School of Medicine, Kyoto University.

Genotyping of three SNPs in the 5' flanking region of MS32

Three SNPs, designated H1, Hf and H2, were reported in the 5' flanking region of MS32 (Monckton et al. 1993). The genotyping and the haplotype analysis of these SNPs were determined according to the protocol established by Monckton et al. (1993). This region (346 bp) was also resequenced in 20 Thai individuals; sequencing reactions were performed using PCR-amplified double-stranded DNA according to the protocol supplied by Perkin Elmer (BigDye Terminator Cycle Sequencing reaction kit™, Applied Biosystems, Foster City, CA). DNAs were analysed on an ABI 3100 system (Applied Biosystems, Foster City, CA). No additional SNPs were discovered.

MVR code mapping of the MS32 alleles

MVR code mapping can be presently approached in three ways: First, by pedigree analysis of diploid codings; second, by mapping of individual separated alleles; and third, allele specific MVR-PCR (Monckton et al. 1993). The first two methods were

omitted from this study for two reasons: 1. the samples we analyzed were obtained on an individual basis and not from pedigrees, and; 2. the amount of genomic DNA obtained was insufficient for individual allele separation.

Therefore, MVR codes of Thai alleles were determined by allele-specific MVR-PCR (Monckton et al. 1993) with minor modifications, using allele-specific PCR primers directed to polymorphic SNP sites in the DNA flanking the minisatellite to amplify a single allele directly from genomic DNA. Two different MVR primers (Jeffreys et al. 1991) are used in separate PCR reactions to amplify two classes of MS32 repeats [**a**-, **t**-type repeats]. PCR products are loaded in adjacent lanes in an agarose gel to generate two complementary ladders, from which the interspersion pattern of repeats can be read for at least 50 repeats into the minisatellite allele; alleles less than 50 repeats long can be mapped in their entirety.

Samples of 10-50 ng of genomic DNA were amplified in 7 µl reactions using the PCR buffer and primers described previously (Jeffreys et al. 1991; Monckton et al. 1993; Jeffreys et al. 1990) plus 0.5 U Taq polymerase. Reactions were cycled for 45 s at 96 °C, and 5 min at 70°C for 19 cycles. PCR products were then electrophoresed through a 35 cm long 1.1% agarose gel (type 1, Sigma-Aldrich) in 0.5 x TBE (44 mM Tris-borate pH 8.3, 1mM EDTA) at 130-140V for 16h, blotted onto Hybond N+ membranes (GE Healthcare), hybridized with a ^{32}P-labelled MS32 repeat probe and visualized by autoradiography.

MS32 allele code database and allele alignment

An MS32 MVR database compiled by AJJ was used to compare Thai and Chinese allele codes with other ethic groups. This database consists of 1072 previously mapped alleles that include 426 north European, 318 Japanese (Nagoya), 2 Han Chinese and 253 African alleles. We added a further 119 Thai alleles plus 71 Han Chinese and 7 Japanese (Nagoya) alleles to give a database of 1269 alleles for the present study. To identify alleles that share regions of map similarity, MVR codes were compared with

each other by heuristic dot matrix analysis using modified Microsoft Excel software originally written by AJJ (Jeffreys et al. 1991; Tamaki et al. 1995). Comparisons searched for perfect 9-repeat matches, and allele pairs showing scores of 22 or more over the best two diagonals (*i.e.* having the greatest allelic similarity) were selected. The authenticity of these selected matches and the final alignment of allele groups were checked by eye, with gaps inserted to improve alignments. Codes for all alleles are available on the authors' Web site (http://www.legal.med.kyoto-u.ac.jp/ms32_database.htm)

Results

SNP frequencies in DNA upstream of MS32 in Thai and Chinese populations

Allelic states of SNPs H1, Hf and H2 used for MVR mapping were determined in 170 Thai individuals. The frequencies of H1G, Hf+ and H2C were 0.80, 0.78 and 0.79, respectively. Genotypes at all three SNPs were in Hardy-Weinberg equilibrium, with ~33% heterozygosity (H) at each site. H1 and Hf frequencies are similar to Japanese, while H2C is more common than in Japanese (H = 0.18) (Monckton et al. 1993). Haplotype analyses of the three SNPs were performed. These SNPs are found in a region of only 302 bp flanking MS32 and show significant linkage disequilibrium ($\chi^2_{[7\ d.f.]} = 101$, $p<0.001$). However, no pair of sites shows complete disequilibrium, as expected given the existence of a meiotic recombination hotspot in this region (Jeffreys et al. 1998). Overall, 109 out of 170 individuals were heterozygous at one or more SNPs, and thus 64% of Thai individuals could have single alleles mapped by allele-specific MVR-PCR, similar to the frequency of informative Japanese. The frequencies observed in the Chinese population were similar to Thai (H1G = 0.84, Hf+ = 0.76, H2C = 0.82), with 64 out of 93 individuals (69%) heterozygous at one or more SNPs. Genotypes at the three SNPs were again in H-W equilibrium, and a similar deviation was found for Chinese between the observed numbers of haplotypes and those expected

at linkage equilibrium ($\chi^2_{[7\ d.f.]} = 57$, $p<0.001$).

MS32 MVR allele code diversity in Thai and Chinese populations

In total, 119 Thai and 73 Han Chinese alleles were mapped by allele-specific MVR-PCR. An example of an MVR mapping autoradiograph is shown in Fig. 1. Typically 50–80 minisatellite repeats could be read from the 5' end of each allele, and thus longer alleles were incompletely mapped. Almost all of the Thai and Chinese alleles mapped were dissimilar (117 out of 119, 72 out of 73, respectively) (Table 1). The sampling distributions of different alleles can be used to estimate allele diversity, $\theta=4Ne\mu$, where *Ne* is the effective population size and μ is the mutation rate. Under the infinite allele model and assuming selective neutrality, θ value is determined from the number of different alleles n_a seen in a sample of *n* individuals by $n_a = \sum_{i=1}^{2n} \{\theta/(\theta+i-1)\}$ and heterozygosity can be estimated as $\theta/(1+\theta)$ (Ewens 1972). The Thai and Chinese data give estimated θ values of 3430 and 2580, respectively. These estimates of diversity suggest an extremely high level of heterozygosity in both populations, at 99.97% and 99.96% respectively. If all alleles are equally rare, then Poisson analysis predicts the existence of more than 1,400 (Thai) and 700 (Chinese) MS32 alleles in order to reach the observed sampling frequency distribution. Since allele frequencies will not be uniform, the true level of allele diversity is likely to be much higher.

Identification of groups of related alleles

In the 7021 possible pairwise dot-matrix comparisons in 119 Thai alleles, 151 comparisons involving 67 alleles had scores of 22 or more, our initial criterion for identifying significantly related alleles. For the 73 Chinese alleles, the 2628 possible pairwise dot-matrix comparisons resulted in 136 comparisons, involving 58 alleles, with scores of at least 22. Visual inspection of the aligned alleles showed a significant sharing of repeat motifs in most cases, with only a few allele pairs being excluded

because they only showed very short motifs scattered along the best diagonals.

We repeated this dot-matrix analysis using a more global database of 1269 alleles, allowing pairs of significantly related alleles to be identified and subsequently assembled into progressively larger groups of alleles sharing common structures. Two of these aligned groups are shown in Fig. 2.

In GroupA Fig.2, 38 alleles form a group. Surprisingly, out of the 1269 possible alleles (which include Caucasians and Africans) this group is composed only of Asian alleles. While the differences in the numbers of samples analyzed for Japanese, Thai and Chinese warrant further assessment, we not only noticed the formation of subgroups, but also that particular alleles such as from Thailand and Papua New Guinea displayed a tendency to group closely together within a subgroup. Incorporation of 5' flanking haplotype data revealed a strong tendency for closely related alleles to share a common haplotype, as expected for divergence from a recent common ancestral allele, and helped to define subgroups; for example alleles 1-9 in Group A (Fig. 2) show a different haplotype from the most of the remaining alleles in this group and presumably represent a distinct sub-lineage. Additional information for other population subsets are available on the authors' Web site (http://www.legal.med.kyoto-u.ac.jp/ms32.htm). The 5' ends of the aligned MVR maps show the most variability, most likely due to the existence of a meiotic recombination hotspot flanking MS32 that appears to be responsible for driving repeat instability (Jeffreys et al. 1988).

Table 2 summarizes the characteristics of allelic similarity between Thai, Chinese, and Japanese populations identified by pairwise comparison with worldwide alleles. For example, 24 of the 119 Thai alleles showed no detectable similarity with any other allele in the database, and a further 32 alleles showing at best only marginal similarities. The remaining 63 alleles (53% of all Thai alleles) showed significant structural relatives and could be classified into 20 groups of related alleles.

Within each group, alleles were ranked in order of allelic similarity as determined by dot-matrix analysis scores and the closest relative of each typed allele from Thais, Chinese and Japanese was identified (Table 2). In the 63 groupable Thai

alleles, 22 (35%) displayed greatest allelic similarity to other Thai alleles, with 4 producing groups composed only of Thais, and 8 belonging to 3 other groups composed only of Asians. The remaining 10 alleles were found in Thai subgroups dispersed throughout 3 larger groups including north Europeans and Africans. 15 of the remaining grouped Thai alleles aligned equally well with Japanese, Chinese and Thai alleles, and 19 other Thai alleles most closely resembled Asian alleles other than Thai, most likely reflecting sampling variation in these small surveys. Within this bracket of 34 alleles, 16 belong to 9 groups composed only of Asians. The remaining 18 contained an assortment of alleles from non-Asian populations that formed Asian subgroups within 4 larger groups composed of various ethnicities. Only 4 alleles were classified as having greatest allelic similarity to north European alleles. These alleles belong to 4 different groups consisting of various ethnicities. None of the alleles were found to have greatest structural similarity to African alleles.

A similar analysis on 73 Chinese alleles showed 6 alleles with no detectable similarity to any other alleles in the database, and a further 24 alleles showing at best only marginal similarities. The remaining 43 alleles (59%) showing significant structural relatives with 6 being most similar to other Chinese alleles and creating two groups of Chinese-only alleles and one group containing other Asian alleles. As with Thais, other Chinese alleles (15 in total) showed equally significant alignments with Chinese, Thai and Japanese alleles and 17 showed preferred alignment with non-Chinese Asian alleles. The remaining 18 Chinese alleles fell into two large groups containing alleles from diverse populations but with clear Asian subgroups. In no case did a Chinese allele preferentially align with an African allele, and in only one case with a north European allele.

The proportion of Thai and Chinese alleles that could be aligned into groups were similar (53% and 59% respectively) but less than for Japanese (69%) ($\chi^2_{[2\ \text{d.f.}]} = 9.2$, $p = 0.010$), in part reflecting different sample sizes with more Japanese alleles having been typed.

Finally, we found 16 unusually short Japanese alleles (Nos. 1-7,11,13-15,17-21;

Fig.2 Group B) within a group of 27 alleles that showed at best only marginal alignments with other alleles. These alleles were all relatively homogeneous with **a**-type repeat arrays terminated in two **t**-type repeats. Alleles with such simple structures could arise by convergent evolution, and thus this group might not have a monophyletic origin. However, most of the short, homogeneous Japanese alleles shared the same G – C haplotype, consistent with monophyly. In contrast, the two short European alleles (Nos.12,16) were on different and distinct haplotypes, suggesting convergence with Japanese allele structures from different ancestral states. The restricted haplotype diversity and relatively high frequency of these short and homogeneous alleles seen in Japanese and to a lesser extent in Thais suggests that these alleles, while structurally not very informative, do mark a distinct lineage largely restricted to east Asia.

Discussion

Extreme variation in minisatellite allele structures provides a potentially powerful tool for analyzing population structures. This is exemplified by global surveys of MVR variation at the MS205 (Armour et al. 1996) and insulin minisatellites (Stead and Jeffreys 2002) that revealed far greater lineage diversity in Africans compared to non-African populations and gave clues about population characteristics during the migration out of Africa. What is not known is whether MVR lineages can be used to probe population structures at a much finer geographical level.

To date, Japan has been the only Asian population to be analysed for diversity in allele structures in minisatellite MS32 (Tamaki et al. 1995). It was unclear whether apparently Japanese-specific groups of related alleles were instead members of allele lineages that are more diffusely spread in Asia. We have started to clarify this issue by extending the analysis to Thai and Chinese alleles, revealing further huge levels of variation in allele structures. Again, there was a strong tendency for groups of the most closely related alleles to show population specificity, either at the level of East Asians in general (Japanese + Thai + Chinese), or more specifically restricted to just one

population as seen for 75% of Japanese alleles but less so for Thai (35%) and Chinese (14%) alleles ($\chi^2_{[2\ d.f.]} = 73.6$, $p<0.0001$). Ignoring alleles showing equally strong alignments over multiple populations and normalising for differing sample sizes showed that an alignable Japanese allele was 10 fold more likely to show closest relationship with another Japanese allele than with a Thai or Chinese allele. This index of population specificity was considerably lower for Thai and Chinese alleles (4-fold, 2-fold, respectively).

Human habitation within closed geographical systems tends to fuel greater allelic specificity. While questions concerning the origin of modern Japanese have a long history of debate, recent evidence suggests the last major admixture of continental populations (*e.g.* Korean, Chinese etc.) to have occurred 13,000 years ago during the Yayoi period. Since then, there have been many sporadic migrational events into Japan from continental Asia. There are many studies comparing mitochondrial DNA polymorphisms in Japanese and surrounding populations (Horai et al. 1996; Tanaka et al. 2004; Tajima et al. 2004). However, maternally inherited mitochondrial DNA has different characteristics in its sex-specific mode of transmission compared to nuclear DNA. This restriction also applies to Y-chromosomes that are transmitted paternally (Hammer and Horai 1995). Additionally, there are a number of phylogenic studies employing STRs (Bowcock et al. 1994; Perez-Lezaun et al. 1997; Chu et al. 1998; Rosenberg et al. 2002; Ayub et al. 2003; Zhivotovsky et al. 2003; Li et al. 2006). The majority of these studies are limited to transcontinental migration and analyses on other ethnicities. One study performs a phylogenic analysis on multiple Asian populations including Japanese using 105 autosomal STR loci and concludes that the Japanese population is more closely related to Southern rather than Northern Chinese (Li et al. 2006). In contrast to these previous studies of multiple STRs, we attempted the present phylogenic analyses using information obtained from only one minisatellite locus. As a result, we not only found that many alleles were population specific, but that loose links could be established between seemingly unrelated ethnic groups (Table 2). Interestingly, while the database contains European and African alleles, the frequency at which the

three Asian ethnic groups displayed the closest relationship to these two groups was extremely low (Japanese 1.4%, Chinese 2.3% and Thai 6.3%), again pointing to strong population specificity of allele lineages and minimal introgression of European alleles into these Asian samples.

The strong ethnic specificity of some lineages of MS32 alleles comes about not only from demographic processes but also from the high rate of mutation, estimated at 1.2% per gamete by pedigree analysis (Jeffreys et al. 1991), coupled with the complex nature of the mutation process. Germline mutations altering array length and MVR pattern are preferentially directed to the 5' end of the repeat array, probably due to the existence of a 5' flanking recombination hotspot (Jeffreys et al 1998). Most sperm mutants at minisatellite MS32 involve inter-allelic conversion-like events involving the copying and pasting of short blocks of repeats from one allele to another (Jeffreys et al 1994). These transfers can be complex, with repeat reshuffling in the transferred segment and with duplication or deletion in the recipient allele at or near the site of transfer. Thus a radically new allele structure can be generated within one or a few mutational steps, creating a new MVR lineage that can then diversify by other more subtle and frequently intra-allelic rearrangements occurring either within or outside the conversion hotspot, as seen at MS32 and other minisatellites [reviewed in Jeffreys et al. (1999)], to create a group of related alleles. Given the high rates of these processes, it is easy to see how a new allele lineage could appear suddenly and recently, and remain restricted to a single population as seen for example in the Japanese. Germline instability also indicates that groups of identical alleles must share a very recent common ancestor, within 125 generations (3000 years) ($p>0.95$) for a pair and even more recently for larger clusters of the same allele. Such very recent divergences predict a strong ethnic specificity for groups of the same alleles, as observed.

Distal to the 5' end of MS32, the incidence of common MVR motifs generally increases, allowing increasingly large groups of related alleles to be constructed. One limitation of the current MS32 MVR procedure is that alleles with more than 50-80 repeats, accounting for 86% alleles, will be incompletely mapped. Some important

lineage information from the 3' ends of alleles will therefore be lost, resulting in single lineages becoming fractured into two or more apparently unrelated sublineages. One challenge will be to develop methods for recovering complete allele structures, as well as algorithms for clustering closely related alleles within the large and structurally diverse lineages that will be so identified to overcome the subjectivity of selecting the most closely related alleles by eye. This has the potential to reveal how different ethnic groups are related and how sub-populations have evolved within the same ethnic group. We found that classifications according to allelic similarity enables not only the identification of major groups (European, African and Asians) but subgroups (e.g. Asian: Thai, Japanese, Chinese) within a larger ethnic group as well. This is made possible by the inherent polarity of mutation in MS32 alleles (Jeffreys et al. 1994), a property not shared by STR loci.

In summary, we have shown that minisatellite MS32 is not only a valuable tool for individual identification but also a potentially powerful system for exploring population structure and evolution at the fine-scale level. Further resolution could be generated by the development of additional autosomal minisatellites as MVR markers, and would complement current mitochondrial DNA and Y-specific markers that can only report on matrilineages and patrilineages respectively.

Acknowledgements This work was supported by grants from the Minister of Education, Science, Sport and Culture of Japan, and from Japan Society for the Promotion of Science.

References

Andreassen R, Olaisen B (1998) De novo mutations and allelic diversity at minisatellite locus D7S22 investigated by allele-specific four-state MVR-PCR analysis. Hum Mol Genet 7:2113-2120

Armour JAL, Harris PC, Jeffreys AJ (1993) Allelic diversity at minisatellite MS205 (D16S309): evidence for polarized variability. Hum Mol Genet 2:1137-1145

Armour JAL, Anttinen T, May CA, Vega EE, Sajantila A, Kidd JR, Kidd KK, Bertranpetit J, Pääbo S, Jeffreys AJ (1996) Minisatellite diversity supports a recent African origin for modern humans. Nat Genet 13:154-160

Ayub Q, Mansoor A, Ismail M, Khaliq S, Mohyuddin A, Hameed A, Mazhar K, Rehman S, Siddiqi S, Papaioannou M, Piazza A, Cavalli-Sforza LL, Mehdi SQ (2003) Reconstruction of human evolutionary tree using polymorphic autosomal microsatellites. Am J Phys Anthropol 122:259–268

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994) High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455–457

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

Chu JY, Huang W, Kuang SQ, Wang JM, Xu JJ, Chu ZT, Yang ZQ, Lin KQ, Li P, Wu M, Geng ZC, Tan CC, Du RF, Jin L (1998) Genetic relationship of populations in China. Proc Natl Acad Sci USA 95:11763–11768

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theor Popul Biol 3:87-112

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

Horai S, Murayama K, Hayasaka K, Matsubayashi S, Hattori Y, Fucharoen G, Harihara S, Park K-S, Omoto K, Pan I-H (1996) mtDNA polymorphism in Asian populations, with special reference to the peopling of Japan. Am J Hum Genet 59:579–590

International Human Genome Sequencing Consortium (2001) Initial sequencing and

analysis of the human genome. Nature 409:860-921

Jeffreys AJ, MacLeod A, Tamaki K, Neil DL, Monckton DG (1991) Minisatellite repeat coding as a digital approach to DNA typing. Nature 354:204-209

Jeffreys AJ, Neumann R, Wilson V (1990) Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. Cell 60:473-85

Jeffreys AJ, Tamaki K, MacLeod A, Monckton DG, Neil DL, Armour JAL (1994) Complex gene conversion events in germline mutation at human minisatellites. Nat Genet 6:136-145

Jeffreys AJ, Murray J, Neumann R (1998) High resolution mapping of cross-overs in human sperm defines a minisatellite-associated recombination hotspot. Mol Cell 2:267-273

Jeffreys AJ, Barber R, Bois P, Buard J, Dubrova YE, Grant G, Hollies CR, May CA, Neumann R, Panayi M, Ritchie AE, Shone AC, Signer E, Stead JD, Tamaki K (1999) Human minisatellites, repeat DNA instability and meiotic recombination. Electrophoresis 20:1665-75

Li SL, Yamamoto T, Yoshimoto T, Uchihi R, Mizutani M, Kurimoto Y, Tokunaga K, Jin F, Katsumata Y, Saitou N (2006) Phylogenetic relationship of the populations within and around Japan using 105 short tandem repeat polymorphic loci. Hum Genet 118:695-707

Monckton DG, Tamaki K, MacLeod A, Neil DL, Jeffreys AJ (1993) Allele-specific MVR-PCR analysis at minisatellite D1S8. Hum Mol Genet 2:513-519

Neil DL, Jeffreys AJ (1993) Digital DNA typing at a second hypervariable locus by minisatellite variant repeat mapping. Hum Mol Genet 2:1129-1135

Pérez-Lezaun A, Calafell F, Mateu E, Comas D, Ruiz-Pacheco R, Betranpetit J (1997) Microsatellite variation and the differentiation of modern humans. Hum Genet 99:1–7

Rosenberg NA, Pritchard JK, Weber JL, Cann HW, Kidd KK, Tajima A, Pan IH, Fucharoen G, Fucharoen S, Matsuo M, Tokunaga K, Juji T, Hayami M, Omoto K,

Horai S (2002) Three major lineages of Asian Y chromosomes: implications for the peopling of east and southeast Asia. Hum Genet 110:80–88

Royle NJ, Clarkson RE, Wong Z, Jeffreys AJ (1988) Clustering of hypervariable minisatellites in the proterminal regions of human autosomes. Genomics 3:352–360

Stead JD, Jeffreys AJ (2002) Structural analysis of insulin minisatellite alleles reveals unusually large differences in diversity between Africans and non-Africans. Am J Hum Genet 71:1273-1284

Tajima A, Hayami M, Tokunaga K, Juji T, Matsuo M, Marzuki S, Omoto K, Horai S (2004) Genetic origins of the Ainu inferred from combined DNA analyses of maternal and paternal lineages. J Hum Genet 49:187–93

Tamaki K, May CA, Dubrova YE, Jeffreys AJ (1999) Extremely complex repeat shuffling during germline mutation at human minisatellite B6.7. Hum Mol Genet 8:879-888

Tamaki K, Huang XL, Yamamoto T, Uchihi R, Nozawa H, Katsumata Y, Jeffreys AJ (1995) Characterisation of MS32 alleles in the Japanese population by MVR-PCR analysis. In: Sawaguchi A. Nakamura S. (eds) DNA polymorphism 3. Toyo Shoten, Tokyo, pp 137-143

Tanaka M., Cabrera V.M., González A.M., Larruga J.M., Takeyasu T., Fuku N., Guo L.J., Hirose R., Fujita Y., Kurata M., Shinoda K., Umetsu K., Yamada Y., Oshida Y., Sato Y., Hattori N., Mizuno Y., Arai Y., Hirose N., Ohta S., Ogawa O., Tanaka Y., Kawamori R., Shamoto-Nagai M., Maruyama W., Shimokata H., Suzuki R., Shimodaira H., Mitochondrial genome variation in eastern Asia and the peopling of Japan,Genome Res. 14 (2004) 1832-1850.

Zhivotovsky LA, Rosenberg NA, Feldman MW (2003) Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. Am J Hum Genet 72:1171–1186

Table 1. Allelic diversity at MS32 analysed by MVR-PCR

		Thai	Chinese	Japanese
No. alleles analyzed		119	73	325
No. times seen in the population	1x	115	71	301
	2x	2	1	5
	3x	-	-	-
	4x	-	-	1
	5x	-	-	2
No. different alleles		117	72	309
θ		3430	2580	3080
Heterozygosity		0.9997	0.9996	0.9997

θ values were estimated by the method of Ewens [12].

Heterozygosity was estimated as $\theta/(1+\theta)$

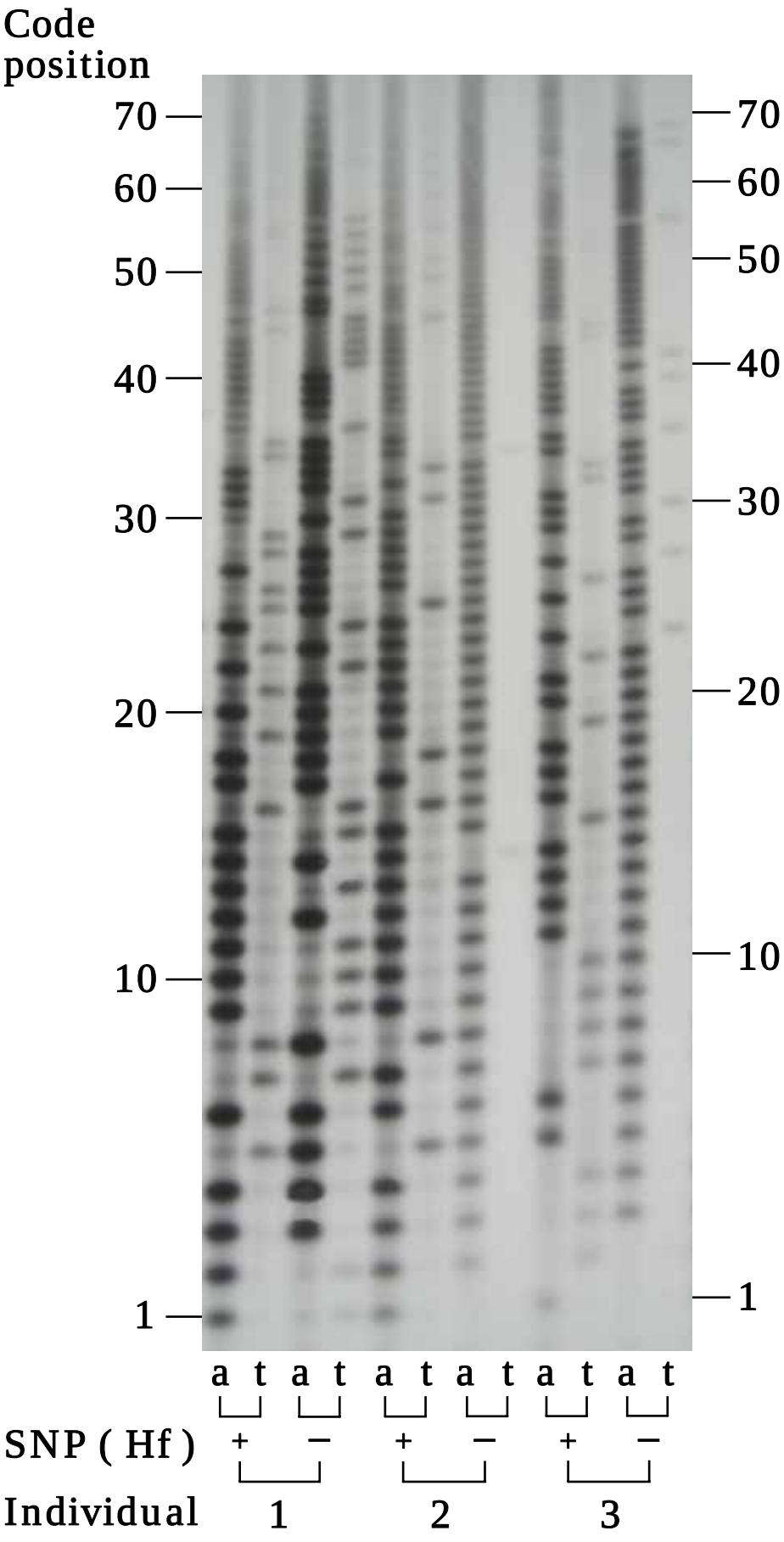
Table 2. Characteristics of Thai, Chinese and Japanese allelic similarity by pairwise comparison with worldwide alleles.

		Thai	Chinese	Japanese
total number of alleles		119	73	325
showed a match (score≥22)		95	67	276
similarity confirmed by eye		63	43	221
greatest allelic similarity to:	same population	22	6	165
	same population plus other Asian populations	15	15	29
	other Asian populations only	19	17	10
	north Europeans	4	1	2
	Africans	0	0	1

Figure Legends

Fig. 1 Examples of allele-specific MVR-PCR. Alleles from the genomic DNA of three individuals were mapped using SNP Hf flanking the minisatellite to amplify single alleles in Hf+/– heterozygotes. MVR-PCR products were visualized after electrophoresis by Southern blot hybridization. MVR codes are read from bottom to top (5’ to 3’) and require no size markers; for example the Hf+ allele in individual 1 can be read as aaaatattaaaaaaat… .

Fig. 2 Examples of alignable MS32 alleles. The ethnic origin R (j, Japanese; ch, Chinese; th, Thai; b, Bangladeshi; png, Papua New Guinean; ne, north European), the haplotype of the three SNPs (H1, Hf and H2) in the 5' flanking DNA and the MVR code of **a**-type, **t**-type and **0**-type repeats are shown for each allele. **0**-type repeats are occasional ‘null’ repeats that fail to amplify in MVR-PCR due to the existence of further (unknown) repeat sequence variants. MVR map segments shared by related alleles are shown in bold, and additional haplotypes shared by some of the grouped alleles are in italics and/or underlined. Gaps (-) have been introduced to improve alignments. Uncertain positions are marked as (?). The ends of short alleles are marked by (<), and the unknown haplotype of long alleles beyond the mapped region are indicated by (…). Alleles tied with “ | ” are indistinguishable. Group B mainly contains alleles consisting of a long succession of **a**-type repeats terminated with two **t**-type repeats at the 3’ end. Due to the largely homogeneous arrays of **a**-type repeats in Group B, with exception of Nos. 22-27, none of the pairwise comparisons could exceed the 22 point threshold and alignments are therefore uncertain.


Code position
70
60
50
40
30
20
10
1
a t a t a t a t a t a t
SNP (Hf) + − + − + −
Individual 1 2 3

Group A

No	R	H1	Hf	H2	MVR
1	j	G	-	C	aaaaaaaaaaaataaaaaaaaaaa-aaaaaaaaaaaataaaataaaaaaaaaaaaaaaaatttatattatttaatt0ataaaatattataat...
2	j	G	-	C	aaaaaaaaaaaaaataaaaaaaaaa--aaaaaaaaaaataaaataaaaaaaaaaaaaaaaatttatattatttaatt0ataaaatatt...
3	ch	G	-	C	ttttaaaaaaaaaaaaaaataaaaaaaaaaaaaaaaaaaaaaataaaataaaaaaaaaaaaaaaaatttatattattt...
4	j		-		aaaaaaaaaaaaataaaaaaaaaa---aaaaaaaaaataaaataaaaaaaaaaaaaaaaatttatattatttaa...
5	j		-		aaaaaaaaaaaaataaaaaaaaaa--aaaaaaaaaaataaaa-aaaaaaaaaaaaaaaatttatatta...
6	j	G	-	C	aaaaaaaaaaaataaaaaaaaaa--aaaaaaaaaaa?aaaataaaaaaaaaaa...
7	j		-		aaaaaaaaaataaaaaaaaaa--aaaaaaaaaaataaaataaaaaaaaaaaaaaaaatttatattatttaatt...
8	j		-		aaaaaaaaaataaaaaaaaaaaaaa--aaaaaaaaaaataaaataaaaaaaaaaaaaaaaatttatatt0tt...
9	j		-		attattaataaaaaaaaaaaaaataaaaaaaaaaaaaaataaaaaaaaaaaaaaaaaatttatattattt...
10	j	G	+	C	?aaaaaaaataaaaataaaataaaaaaa-aaaaaaaatttatattatttaattaataa-aata...
11	th	G	+	C	atatttaataatatttaataaaaaaa-aaaaaaaatttatattatttaattaataa-aata-ataaa...
12	th	G	+	C	?a?aaatttaataaaaaaa-aaaaaaaat0tatattatt?aa...
13	th	G	+	C	ttaaaaaatttatattatttaattaataa-aata-ataaaattaaaaaaataaaatttaaataaa...
14	ch	G	+	C	aaaataaaaataaaataaaaaaaaaataaaataaaaaaa-aaaaaaaatttatattatttaattaataa-aata-attaataaa...
15	ch	G	+	C	aataaaaaaaa0aaaaaaaaaaaaaaataaaaaaaaaaa-aaaaaaaatttatattatttaattaataa-aaaaaaaa...
16	ch	G	-	T	aaaaaaaaaaaaaaaaaaaaaaataaatataaaataaaataaaaaaa--aaaaaaatttatattatttaattaataa-aata-ataaatttata...
17	ch	G	+	C	taattaaaaaaaaaaaaaaaaaaaaaataaaaaaaaaaaaaaaattttttatattatttaattaataa-aata-ataaaattaaaa...
18	j		+		aaaaaaataaaataaaaaaa-aaaaaaaatttatattatttaattaataa-aata-ataaaattaaaaaataaaatttaaaataaaaaaaa...
19	j	G	-	C	aaaaaaaaataaaataaaaaaaaaaaaaaaaatttatattatttaattaataa-aata-ataaaattaaaat...
20	j	G	+	C	aaaaaaaaataaaataaaaaaaaaaaaaa-----atattatttaattaataa-aata-ataaaattaaaaaataaa...
21	j	G	+	C	taaaaaattttaaaaaaaaaaaaaaaattaaaaaaaaatttatattatttaaa-aataa-aata-ataaaattaaaaa...
22	j	G	+	C	aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaatatttaattaataa-aata-ataaaattaaaaaat...
23	j	C	+	C	aataaaaataaaaaaatttattatttaattaataa-aata-ataaaattaaaataaaaaaaattttaaa...
24	j	G	+	C	aaaaatttaaaaaaaaaaatttattatttaattaataa-0ata-ataaaattaaaataaaaaaaattttaaat...
25	j	G	+	C	aaaaaaaaatttattatttaattaataa-0ata-ataaaattaaaataaaaaaaattt...
26	j		+		aaaaaaaatttattatttaattaataa-0ata-ataaaattaaaataaaaaaaattttaaataaaaa...
27	j		+		aaaaaaaaa0aataaaataaaaaaaaaaaaaaaaaatttatattatttaaataataaaaata-ataaaatta----ataaaatttaaaa...
28	j		+		aaaaaaaaaaaaaaaaaaaaaaaatttatattatttaaataataaaataa-attaaaata----ataaaatttaaaataaaaaa...
29	j	G	+	T	taaaaaaaaaaaaaaaaaaatttaatatatttaaataataaaataaaattaaaata----ataaaatttaaaataaaaaaaaa...
30	th	G	+	C	atataaaataaaataaaattaaaata----ataaaatttaaaataaaaaaaaaaaataa...
31	j	G	+	C	taaaaaaaaaaaaaaaaatttatattatttaaataataaaataaaattaaaata----ataaaatttaaaataaaaaaaaa0aat0attaattt...
32	b				???aaattatttaaaaaaatataaaaataaa-attaaaata----ataaaatttaaaataaaaaaaaaaaaa...
33	j	G	+	C	aaatttttttaattaataaaataataaa-attaaaata----ataaaatttaaaataaaaaaaaa0aat0attaaattaattaattataataaaa...
34	j	G	+	C	tt0aataaaataataaa-attaaaata----ataaaatttaaaataaaaaaaaa0aat0attaaattaattaattataataaaatt.
35	png		-		ttaaaaaaatttatattatttaattaataaaatattatttaattaataa-aata-ataaaattaa-ataata...
36	png		+		aaatttatataaaaatttatattatttaattaataa-aata-ataaaattaaaataatataaaaaaaaaaa...
37	png		+		?ataaaaatttatattatttaatta-taa-aata-ataaaattaaaataatataaaaaaaaaaaaaaaaaat...
38	png		+		aata-ataaaattaaaataatataaaaaaaaaa-aaaaaaaattttaattatataaaaaaataattaaaaa....

Group B

No	R	H1	Hf	H2	MVR
1	j	G	-	C	ataaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa...
2	j		-		ataaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaatt<
3	j		-		tataaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaatt<
4	j		-		aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaatt<
5	j	G	-	C	ataataaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaatt<
6	j		-		ataaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaatt<
7	j	G	-	C	ataaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaatt<
8	th	G	-	C	aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaatt<
9	th	G	-	C	?????aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaatt<
10	th	G	-	T	aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaatt<
11	j	G	-	C	| aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaatt<
12	ne	G	+	T	| aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaatt<
13	j	G	-	C	aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaatt<
14	j	G	-	C	aaaaaaaaaaaaaaaaaaaaaaaaaaaaatt<
15	j	G	-	C	aaaaaaaaaaaaaaaaaaaaaaaaatt<
16	ne	G	+	C	aaaaaaaaaaaaaaaaaaaaaaaatt<
17	j		-		aaaaaaaaaaaaaaaaaaaaaaatt<
18	j	G	-	C	aaaaaaaaaaaaaaaaaaaaaatt<
19	j		-		aaaaaaaaaaaaaaaaaaatt<
20	j	G	-	C	aaaaaaaaaaaaaaaaatt<
21	j	G	-	C	aaaaaaaaaaaaaatt<
22	j		+		aaaaaaaaaaaaaaaaaaaaaaaaaaaaaataaaaaataaataaaaat--aaaaaaaaaaaaaaaaaaaaaaaaaaaatt<
23	th	C	+	C	aaaaaaaaaaaaaataaaaaataaataaaaat--aaaaaaaaaaaaaaaaaaaaaaaaaaaatt<
24	j	C	+	C	taaaaaaaaaaaaaaataaaaaataaataaaaaaataaaaaaaaaaaaaaaaaaaaaaaaaaaatt<
25	j	C	+	C	aaaaattaaaaaaaaaaaaaataaaaaataaataaaaat--aaaaaaaaaaaaaaaaaaaaaaaaa......
26	th	G	+	T	atataaaaaaaaaaaaaaaaaa?aataaaaaataaataaaaataaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaatt<
27	j	C	+	C	aaataaaataaaaaataaataaaaat---aaaaaaaaaaaaaaaaaaaaaaaaaatt<