ORIGINAL ARTICLE

B. Budowle · K. L. Monson · R. Chakraborty

Estimating minimum allele frequencies for DNA profile frequency estimates for PCR-based loci

Received: 7 June 1995 / Received in revised form: 7 August 1995

Abstract In order that there can be confidence that DNA profile frequency estimates will not place undue bias against a defendant, 2 methods are described for estimating minimum allele frequency bounds for PCR-based loci. One approach estimates minimum allele frequencies for VNTR and STR loci using sample size and the observed heterozygosity at a locus, while the second approach, appropriate for loci typed with allele-specific oligonucleotide probes, is based only on sample size. The use of a minimum allele frequency enables compensation for sparse sampling of infrequent alleles in population databases.

Key words Allele Frequency · VNTR · STR · PM Loci · Population Data

Introduction

Highly polymorphic loci, such as minisatellites and microsatellites, can contain a large number of alleles and many alleles could be rare. To allay concerns that estimates presented in legal cases might underestimate the frequency of occurrence of DNA profiles, and therefore place undue bias against a defendant, some procedure should be used to compensate for sparse sampling of infrequent alleles in population databases. For the variable number of tandem repeat (VNTR) loci detected by restriction fragment length polymorphism (RFLP) typing, many in the forensic community in North America use the fixed bin method to classify quasi-continuous distributions of alleles. To allow for use of small-size databases and provide a bound on rare allele frequencies a minimum allele frequency of rare alleles is derived by a process termed "rebinning" whereby bins with fewer than 5 counts are merged with contiguous bins (Budowle et al. 1991 b).

For polymerase chain reaction (PCR)-based loci more discrete allelic data can be obtained than is possible with VNTRs typed by RFLP analysis. Some VNTR loci, such as D1S80 (Kasai et al. 1990; Budowle et al. 1991 a) and ApoB (Boerwinkle et al. 1989; Ludwig et al. 1989), can be amplified first by PCR and the amplified products subsequently resolved by electrophoresis into allele classes based on the size of the repeat. The short tandem repeat (STR) loci are a subgroup of these VNTR loci. These loci are highly polymorphic and are abundant in the human genome (Edwards et al. 1991, 1992). Because the allele size of STRs is generally less than 350 base pairs, they are amenable to amplification by the PCR (Saiki et al. 1985; Edwards et al. 1991). The STR loci are composed of tandemly repeated sequences 2–5 base pairs in length. The amplified products of STR loci can be resolved to at least one repeat unit by separation on denatured polyacrylamide gels (Edwards et al. 1991).

DNA from forensic samples also can be typed with loci whose alleles are due to variation in their nucleotide sequence. Currently, 6 loci can be amplified simultaneously and typed: HLA-DQα (Gyllensten and Erlich 1988; Saiki et al. 1989), low density lipoprotein receptor (LDLR) (Yamamoto et al. 1984), glycophorin A (GYPA) (Siebert and Fukuda 1987), hemoglobin G gammaglobin (HBGG) (Slightom et al. 1980), D7S8 (Horn et al. 1990), and group-specific component (Gc) (Yang et al. 1985) [also known collectively as PM loci; Amplitype PM DNA Test System, Roche Molecular Systems, Alameda, Calif; Herrin et al. (1994)].

This paper implements 2 statistical/population genetics approaches for estimating a minimum allele frequency for

This is publication number 95–10 of the Laboratory Division of the Federal Bureau of Investigation. Names of commercial manufacturers are provided for identification only, and inclusion does not imply endorsement by the Federal Bureau of Investigation

B. Bundowle (✉) · K. L. Monson
Forensic Science Research and Training Center, FBI Laboratory, FBI Academy, Quantico, Virginia 22135, USA

R. Chakraborty
Center for Demographic and Population Genetics, University of Texas School of Biomedical Sciences, Houston, Texas 77225, USA

each of the PCR-based loci. The methods offer a better-founded approach than ad hoc minimum allele frequency bounds, such as 0.05 (proffered by NRC Report 1992; I. Evett and R. Fourney personal communications). The first method, under the predictions of the Infinite Allele Model (IAM) (Kimura and Crow 1964; Ewens 1972), defines a minimum allele frequency for the VNTR and STR loci, which can be estimated from the sample size and the observed heterozygosity in a database sample. The second approach, based solely upon sample size, is appropriate for the PM loci where the total number of observed alleles is fixed by the assay (and could be applied to VNTR and STR loci, as well).

Materials and methods

The databases used in this study, as examples, were reported previously (Budowle et al. 1995a, b; Hochmeister et al. 1994; Huang et al. 1994, 1995; Kloosterman et al. 1993).

The theory described by Chakraborty (1992) was used to determine the minimum allele frequency for VNTR/STR loci. The minimum allele frequency for a database is estimated with 100 (1 – α) percent confidence by

$$p_{\min} = 1 - [1 - (1 - \alpha)^{\frac{1}{c}}]^{\frac{1}{2n}}$$

where $p_{\min}$ is the minimum allele frequency, c is the number of common alleles which can be estimated from the level of heterozygosity, and n is the number of individuals (see Chakraborty 1992 for derivation and effectuation). A rare allele was defined as any allele that occurs less frequently than 0.01 in a sample database (Nei 1975; Neel 1973; Chakraborty 1981). In this study α was set at 0.05.

The number of alleles for PM loci is fixed by the number of probes immobilized on the nylon strips. The above described approach, founded on the IAM, would therefore not be applicable.

The minimum allele frequency for a database was estimated according to the suggestion by Weir (1992) per Nelson (1978) as

$$p_{\min} = 1 - \alpha^{\frac{1}{2n}}$$

where the variables are defined as before.

Table 1 Minimum allele frequency estimates for VNTR and STR loci (α = 0.05)

Population/Locus	Sample Size (N)	Het.[a]	Minimum Frequency[b]	Minimum Frequency[c]
Chinese/HUMTHO1[d]	116	0.681	0.0220	0.0128
Chinese/TPOX[d]	116	0.621	0.0212	0.0128
Chinese/CSF1PO[d]	116	0.698	0.0222	0.0128
Swiss/VWA[e]	100	0.820	0.0278	0.0149
Swiss/HUMTHO1[e]	100	0.820	0.0278	0.0149
Swiss/F13A1[e]	99	0.768	0.0276	0.0150
African American/ D1S80[f]	606	0.870	0.0049	0.0025
US Caucasian/D1S80[f]	718	0.784	0.0039	0.0021
Chinese/D1S80[g]	105	0.905	0.0287	0.0142
Dutch/D1S80[h]	150	0.790	0.0183	0.0099
SE Hispanics/D1S80[f]	247	0.806	0.0114	0.0060
SW Hispanics/D1S80[f]	162	0.796	0.0171	0.0092

[a] Het = Heterozygosity
[b] Chakrabory (1992)
[c] Nelson (1978)
[d] Huang et al. (1995)
[e] Hochmeister et al. (1994)
[f] Budowle et al. (1995a)
[g] Huang et al. (1994)
[h] Kloosterman et al. (1993)

Table 2 Minimum allele frequency estimates for PM loci (α = 0.05)

Population/Locus	Sample Size (N)	Minimum Frequency[a]
African American/PM and DQA1 Loci[b]	145	0.0103
US Caucasian/PM and DQA1 Loci[b]	148	0.0101
SE Hispanic/PM and DQA1 Loci[b]	94	0.0158
SW Hispanic/PM and DQA1 Loci[b]	96	0.0155

[a] Nelson (1978)
[b] Budowle et al. (1995b)

Results and discussion

Tables 1 and 2 display examples of minimum allele frequencies for a number of VNTR, STR, and PM loci. The estimates (α = 0.05) are based on heterozygosity (for VNTR and STR loci) and/or sample size (for PM loci) instead of ad hoc approaches. The 2 methods do not yield substantially different minimum allele frequencies, with the Chakraborty (1992) approach yielding larger minimum allele frequency estimates for VNTR and STR loci. However, the Chakroborty (1992) approach is a more desirable estimator of minimum allele frequencies for VNTR and STR loci, since it incorporates more information, i.e., both sample size and heterozygosity.

Obviously, as sample size increases, greater confidence in allele frequency estimates can be obtained and the minimum allele frequency bound decreases. When sample size is held constant, for the Chakraborty (1992) approach, and heterozygosity increases the minimum allele frequency also increases – the minimum allele frequency must increase to account for decreased precision in the observed frequency of alleles that are rare for a given database. Therefore, minimum frequency estimates of DNA profiles containing rare alleles for the more polymorphic loci will tend to be larger than for less polymorphic loci.

For the Chakraborty (1992) approach, a rare allele was defined as any allele with a frequency less than 0.01. A rare allele could have been defined with a higher minimum allele frequency. However, according to theory as the bound for a rare allele increases, there are fewer common alleles (c), and the minimum allele frequency estimate decreases. Therefore, a designation of a rare allele based on a fequency of 0.01 is appropriate.

One could argue, that instead of the IAM, an alternative mutation model for the generation of new alleles at a locus, i.e. the step-wise mutation model (SMM), may be more appropriate for some VNTR and STR loci because the generation of new alleles may be subject to different mutational forces (Shriver et al. 1993; Valdes et al. 1993). While this may be true, there is little concern for forensic applications. When the mutation rate is considered the same, the IAM prediction of heterozygosity is higher than that of the SMM, and hence for the same size database, the IAM generates larger minimum allele frequencies than the SMM (Ohta and Kimura 1973).

Since the number of alleles for the PM loci is predetermined and fixed by the nubmer of sequence-specific oligonucleotide probes used, neither the IAM nor SMM is appropriate for determining the expected number of alleles to be observed in a sample database at such loci. Consequently, information on heterozygosity cannot be utilized in estimating the minimum allele frequencies for PM loci. Thus, the minimum allele frequency for PM loci is based only on sample size (Table 2). These minimum bounds also generate a threshold for rare alleles that should yield a conservative profile frequency, when rare alleles are part of the DNA profile.

In conclusion, minimum allele frequencies for PCR-based loci, based on statistical and population genetics theory, were determined in order that there can be confidence that DNA profile frequency estimates are meaningful even with small size databases. Caution should be taken not to infer similarities or dissimilarities among different databases based on the tabulated minimum allele frequencies. If databases differ in sample size, and/or in heterozygosity, minimum allele frequencies will differ. Therefore, DNA profile frequency comparisons between databases, particularly where minimum allele frequencies are invoked, may not be similar even if the samples are derived from the same ethnic group.

References

Boerwinkle E, Xiong W, Fourest E, Chan L (1989) Rapid typing of tandemly repeated hypervariable loci by the polymerase chain reaction: application to the apolipoprotein B 3′ hypervariable region. Proc Natl Acad Sci USA 86:212–216

Budowle B, Chakraborty R, Giusti AM, Eisenberg AJ, Allen RC (1991 a) Analysis of the variable number of tandem repeats locus D1S80 by the polymerase chain reaction followed by high resolution polyacrylamide gel electrophoresis. Am J Hum Genet 48:137–144

Budowle B, Giusti AM, Waye JS, Baechtel FS, Fourney RM, Adams DE, Presley LA, Deadman HA, Monson KL (1991 b) Fixed bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci for use in forensic comparisons. Am J Hum Genet 48:841–855

Budowle B, Baechtel FS, Smerick JB, Presly KW, Giusti AM, Parsons G, Alevy M, Chakraborty R (1995 a) D1S80 population data in African Americans, Caucasians, Southeastern Hispanics, Southwestern Hispanics, and Orientals. J Forensic Sci 40(1): 38–44

Budowle B, Lindsey JA, DeCou JA, Koons BW, Giusti AM, Comey CT (1995 b) Validation and population studies of the loci LDLR, GYPA, HBGG, D7S8, and Gc (PM loci), and HLA-DQα using a mulitplex amplification and typing procedure. J Forensic Sci 40:45–54

Chakraborty R (1981) Expected number of alleles per locus in a sample and estimation of mutation rates. Am J Hum Genet 33: 481–484

Chakraborty R (1992) Sample size requirements for addressing the population genetic issues of forensic use of DNA typing. Hum Biology 64:141–159

Edwards A, Civitello A, Hammond HA, Caskey CT (1991) DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. J Hum Genet 49:746–756

Edwards A, Hammond H, Jin L, Caskey CT, Chakraborty R (1992) Genetic variation at five trimeric and tetrameric repeat loci in four human population groups. Genomics 12:241–253

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

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

Herring G, Fildes N, Reynolds R (1994) Evaluation of the AmpliType PM DNA test system on forensic case samples. J Forensic Sci 39:1247–1253

Hochmeister MN, Jung MM, Budowle B, Borer UV, Dirnhofer R (1994) Swiss population data on three tetrameric short tandem repeat loci – VWA, HUMTHO1, and F13A1 – derived using multiplex PCR and laser fluorescence detection. Int J Legal Med 107:34–36

Horn GT, Richards B, Merrill JJ, Klinger KW (1990) Characterization and rapid diagnostic analysis of DNA polymorphisms closely linked to the cystic fibrosis locus. Clin Chem 36:1614–1619

Huang NE, Chakraborty R, Budowle B (1994) D1S80 allele frequencies in a Chinese population. Int J Legal Med 107:118–120

Huang NE, Schumm J, Budowle B (1995) Chinese population data on three tetrameric short tandem repeat loci – HUMTHO1, TPOX, and CSF1PO – derived using multiplex PCR and manual typing. Forensic Sci Int 71:131–136

Kasai K, Nakamura Y, White R (1990) Amplification of a variable number of tandem repeat (VNTR) locus (pMCT118) by the polymerase chain reaction (PCR) and its application to forensic science. J Forensic Sci 35:1196–1200

Kimura M, Crow JF (1964) The number of alleles that can be maintained in a finite population. Genetics 49:725–738

Kloosterman AD, Budowle B, Daselaar P (1993) PCR-amplification and detection of the human D1S80 VNTR locus: Amplification conditions, population genetics, and application in forensic analysis. Int J Legal Med 105:257–264

Ludwig EH, Friedl W, McCarthy BJ (1989) High resolution of a hypervariable region in the human apolipoprotein B gene. Am J Hum Genet 45:458–464

National Research Council (1992) DNA typing: statistical bases for interpretation. In: DNA Technology in Forensic Science, Chapter 3. National Academy Press, Washington DC, pp 74–96

Neel JV (1973) 'Private' genetic variants and the frequency of mutations among South American Indians. Proc Natl Acad Sci USA 70:3311–3315

Nei M (1975) Molecular population genetics and evolution. North Holland/American Elsevier. Amsterdam New York, p 118

Nelson SL (1978) Nomograph for samples having zero defectives. J Qual Technol 10:42–43

Ohta T, Kimura M (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet Res 22:201–204

Saiki RK, Scharf S, Faloona T, Mullis KB, Horn GT, Erlich HA, Arnheim N (1985) Enzymatic amplification of beta-globin genomic sequences and restriction analysis for diagnosis of sickle cell anemia. Science 230:1350–1354

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

Shriver MD, Jin L, Chakraborty R, Boerwinkle E (1993) VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. Genetics 134:983–993

Siebert PD, Fukuda M (1987) Molecular cloning of human glycophorin B cDNA: nucleotide sequence and genomic relationship to glycophorin A. Proc Natl Acad Sci USA 84:6735–6739

Slightom JL, Blechl AE, Smithies O (1980) Human fetal $^{G}\gamma$- and $^{A}\gamma$- globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell 21:627–638

Valdes AM, Slatkin M, Freimer NB (1993) Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics 133:737–749

Weir BS (1992) Independence of VNTR alleles defined by fixed bins. Genetics 130:873–887

Yamamoto T, Davis CG, Brown MS, Schneider WJ, Casey ML, Goldstein JL, Russell DW (1984) The human LDL receptor: a cysteine-rich protein with multiple Alu sequences in its mRNA. Cell 39:27–38

Yang F, Brune JL, Naylor SL, Apples RL, Naberhaus KH (1985) .Human group-specific component (Gc) is a member of the albumin family. Proc Natl Acad Sci USA 82:7994–7998