International Congress Series 1288 (2006) 495–497

www.ics-elsevier.com

An assessment of the behavior of the population structure parameter, θ, at the CODIS loci

A.D. Anderson, B.S. Weir *

Program in Statistical Genetics, Department of Statistics, North Carolina State University, USA

Abstract. The population structure parameter, θ, is important in the calculation of forensic DNA match probabilities. In this work, we use the Phase I HapMap dataset to examine the behavior of θ in chromosomal regions containing CODIS loci. This data supports the premise that the same value of θ can be used for all CODIS loci, but suggests that some CODIS loci, most noticeably vWA and D13S317, may occur in genomic regions where natural selection may be causing violations from the usually adopted neutral model.

Keywords: F_{ST}; HapMap; SNP data; Population structure

1. Introduction

When assessing the significance of matching and partially matching DNA profiles, probability calculations are made under a model that includes a population structure parameter, θ (also known as F_{ST}), that allows for correlations between alleles found in members of the same subpopulation of the population from which allele frequencies were estimated [1–4]. This model relies on several assumptions, such as (1) the degree of population structure is constant over loci (i.e. the same value of θ may be used at all loci) and (2) the loci in question are neutral. This first assumption may be violated because different portions of the genome have somewhat different genealogical histories, and, indeed, recent research [5] has shown that θ estimates vary considerably throughout the genome. The second of these assumptions has to do with whether a neutral model with population structure parameter θ is sufficient to capture the variation in allele frequencies among the subpopulations. For example, if a locus is under selection, an advantageous

* Corresponding author. Tel.: +1 919 515 3574.
E-mail address: weir@stat.ncsu.edu (B.S. Weir).

0531-5131/
doi:10.1016/j.ics.2005.10.040

allele may arise and sweep to prominence in a subpopulation, causing the allele frequencies in that subpopulation to be more different than expected under the neutral model. This effect would carry over, through gene hitch-hiking, to any locus sufficiently close to the locus under selection.

In this work, we used a high-density SNP data set [6] to investigate the degree of population structure near the CODIS loci in an attempt to determine whether it appears as if the above two assumptions hold.

2. Materials and methods

We used SNP data on unrelated individuals from the Phase I HapMap [6] dataset, restricting our attention to those SNPs that were polymorphic in all subpopulations. The HapMap data set we used had about 600,000 markers genotyped on 60 Caucasians of European descent (CEU), 60 Yoruba from Ibadan, Nigeria (YRI), 45 Han Chinese from Beijing (CHB) and 44 Japanese from Tokyo (JPT).

Two methods were used to estimate the degree of population structure among these groups. The first method [7] assumes a constant value of θ for all subpopulations and so is used to estimate a population-average value of θ. The second method [8] allows for θ to vary between the subpopulations, resulting in a different population-specific estimate for each population. For both methods, preliminary results showed tremendous variability in

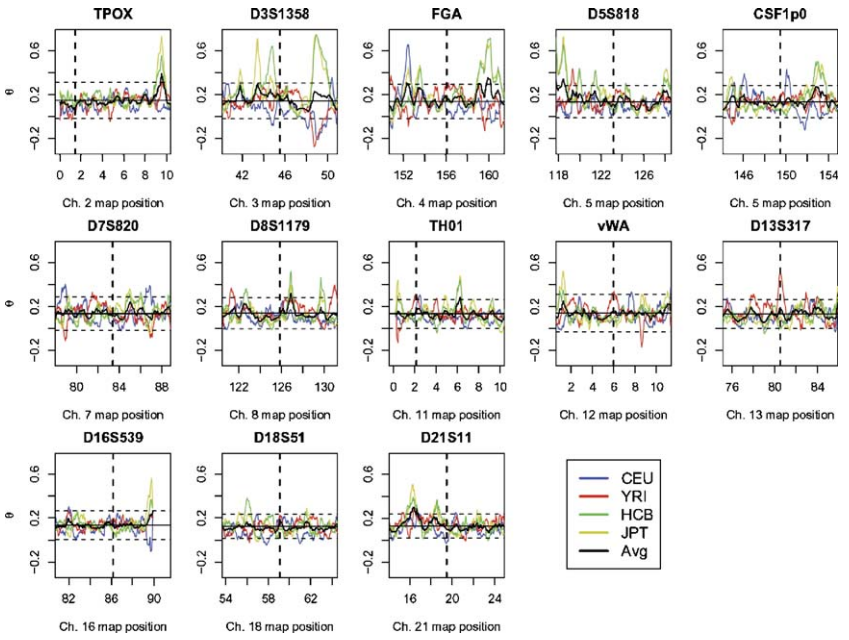

Fig. 1. Population structure estimates in regions containing the CODIS loci. The vertical dashed line in each plot represents the location of the CODIS locus. The horizontal lines represent the mean and mean $\pm$ 3 standard deviation boundaries of the population-average estimates of θ across the chromosome. In all plots, the map position is measured in Mb.

single-locus estimates, even for nearby loci, so, in an attempt at smoothing the estimates, we chose to estimate θ using all loci within a sliding 500 kb window.

3. Results and discussion

One assumption we wanted to evaluate was whether or not it was appropriate to use the same value of θ for all CODIS loci. Since (1) the CODIS loci are microsatellites and our θ estimates are based on SNPs and (2) our data set consisted of subpopulations of the entire human species, whereas the θ values used in forensic calculations refer to subpopulations of specific ethnic groups, our results cannot be used to give estimates of possible values of θ that might be used. Nevertheless, if SNP-based θ estimates vary widely, it can be assumed that nearby microsatellites will also vary, so this data does provide an indication of how θ behaves near the CODIS loci. For evaluating whether a single θ estimate is appropriate, the population-average θ estimate is of primary interest because, when making forensic probability calculations we do not specify the subpopulation to which the suspect(s) and perpetrator(s) may belong. Based on the results shown in Fig. 1, the population-specific values are quite constant across the loci.

We next looked for evidence of selection. In this case, the population-specific θ estimates are of greater interest, because one signature of selection is that certain subpopulations will have allele frequencies that vary considerably from the other subpopulations. The population-average estimate, which is based on a supposition of some degree of homogeneity among the subpopulations, can obscure this effect. Fig. 1 reveals that a few of the CODIS loci, most noticeably vWA and D13S317, are located near peaks in the population-specific θ estimates for one or more subpopulations. This raises some concern about these loci and should certainly be examined in more depth.

Acknowledgements

This work was supported in part by NIH grants GM 45344 and ES 7329.

References

[1] D.J. Balding, R.A. Nichols, DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands, Forensic Sci. Int. 64 (1994) 125–140.
[2] B.S. Weir, The effects of inbreeding on forensic calculations, Annu. Rev. Genet. 28 (1994) 597–621.
[3] D.J. Balding, R.A. Nichols, Significant genetic correlations among Caucasians at forensic DNA loci, Heredity 78 (1997) 583–589.
[4] B.S. Weir, Matching and partially-matching DNA profiles, J. Forensic Sci. 49 (2004) 1009–1014.
[5] B.S. Weir, et al., Measures of human population structure show heterogeneity among genomic regions, Genome Res. 15 (2005) 1468–1476.
[6] International HapMap Consortium, A haplotype map of the human genome, Nature 437 (2005) 1299–1320.
[7] B.S. Weir, C.C. Cockerham, Estimating *F*-statistics for the analysis of population structure, Evolution 38 (1984) 1358–1370.
[8] B.S. Weir, W.G. Hill, Estimating *F*-statistics, Annu. Rev. Genet. 36 (2002) 721–750.