International Congress Series 1261 (2004) 88–90

www.ics-elsevier.com

How much more should the Y-STR Haplotype Reference Database increase to reach a pragmatic saturation level?

Luisa Pereira[a,*], António Amorim[a,b]

[a] *IPATIMUP (Instituto de Patologia e Imunologia Molecular da Universidade do Porto), Rua Dr. Roberto Frias s/n 4200-465, Oporto, Portugal*
[b] *Faculdade de Ciências da Universidade do Porto, Portugal*

Abstract. Empirical saturation curves can be a valuable tool for estimation of sampling effort required to detect an adequate number of Y-STR haplotypes, since they are independent of theoretical models. Nevertheless, they were shown to be highly dependent on current sample size, as detected when applied to the study of the Y-STR Haplotype Reference Database, at a European scale. The influence of geography on sampling saturation was difficult to evaluate due to high heterogeneity of the sample sizes in the database, but at the European level, it is probably very slight. Concerning the present status of the database, it is expected that a fivefold increase in the sample size of the database (up to 60,000 individuals) will reveal twice the current number of haplotypes (11,308 haplotypes), reaching a pragmatic saturation level of 5% (that is, when along with a sample increment of M individuals, no more than $0.05 \times M$ "new" haplotypes are added).

Keywords: Y-STR haplotypes; Saturation curves; Saturation levels; Database

1. Introduction

Statistical problems, namely sampling strategic, when dealing with hypervariable loci such as mtDNA sequences or Y-chromosome STR haplotypes are particularly acute. Among these problems of notorious practical interest are those related to frequency estimations (for instance in evaluating genetic relationships between populations) and to the so-called sampling saturation (i.e., the sample size required for detecting essentially all the distinct items under study in the population). It has been shown, extrapolating from actual European mtDNA HVRI sequence data, that current sample sizes are clearly insufficient for adequate population comparisons, requiring a three- to fourfold increase of the sampling effort per population [1]. In accordance, when analysing the same type of data in an unusually large ($n = 1200$) German sample registered that although haplotype diversity values reached a plateau around $n = 400$, the number of different haplotypes continued to steadily increase [2].

* Corresponding author. Tel.: +351-22-557-0700; fax: +351-22-557-0799.
E-mail address: lpereira@ipatimup.pt (L. Pereira).

0531-5131/
doi:10.1016/S0531-5131(03)01653-4

In this work, we propose to estimate the saturation levels for the number of Y-STR haplotypes in Europe, as currently reported in the Y-STR Haplotype Reference Database, by using empirical saturation curves, which are independent of theoretical models.

2. Material and methods

The Y-STR Haplotype Reference Database was downloaded on 05/04/2003, and from the 12,802 minimal haplotypes present there, we have used the ones for populations in Europe (excluding the ethnic group Romani), summing up 11,692 individuals corresponding to 5524 haplotypes; Y-STR loci of the minimal core (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393 and DYS385) were considered for analysis. The strategy used consists in performing simulations, which generate sets of random subsamples (four independent assays, by using the function RAND( ) of dBase III+ program) of variable size in sufficient numbers for a direct curve fitting approach. Saturation curves were obtained in the CurveExpert version 1.37 package (http://www.curveexpert.webhop.biz/), under the saturation growth-rate model ($y=(ax)/(b+x)$). Pragmatic saturation levels were set at: 5%, when along with an increment of M individuals, no more than $0.05 \times M$ "new" haplotypes are added; and 1%, when along with an increment of M individuals, no more than $0.01 \times M$ "new" haplotypes are added.

3. Results

The saturation curves obtained for the number of Y-STR haplotypes in each population showed a high heterogeneity inside each geographic area (e.g., central–eastern European populations represented in Fig. 1).

Interestingly, curves with the highest saturation values were observed for populations in which the current sample size is bigger. This indicates that current sample size is a main factor for empirical sampling saturation evaluation.

We then evaluated whether geography had any influence on the sample saturation estimates. We compared the saturation curves obtained in the two ways: (a) each point corresponding to the observed values for each country (and one for the total European

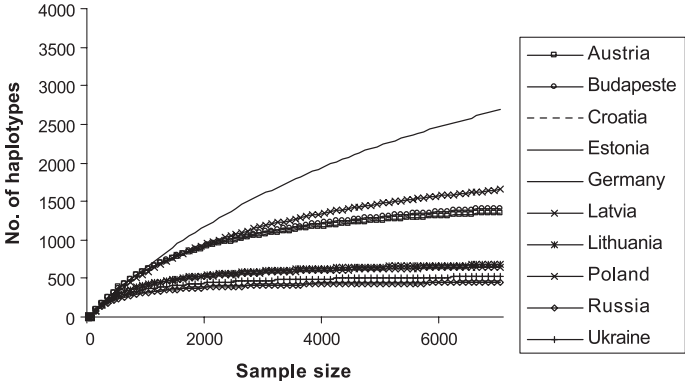

Fig. 1. Saturation curves for the number of Y-STR haplotypes in some European populations.

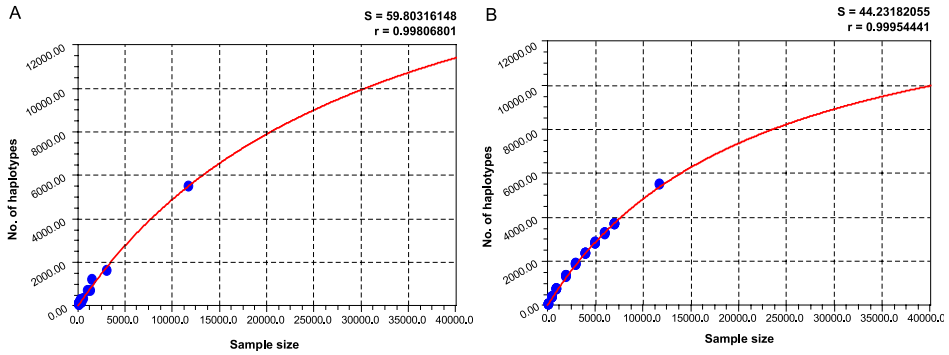

Fig. 2. Saturation curves for the number of Y-STR haplotypes without (A) and with (B) geographic randomisation, as described in the text.

sample; Fig. 2a); (b) using the European sample for simulating points of various sample sizes in four randomised assays (Fig. 2b). There was no significant increase in the correlation coefficient ($r = 0.99806801$ and $r = 0.99954441$, respectively) with geographic randomisation, suggesting a minor geographic influence on the saturation values for the number of Y-STR haplotypes at the European level.

The saturation values obtained with the Y-STR Haplotype Reference Database were: (a) 60,000 individuals corresponding to 11,308 haplotypes, at the 5% level; and (b) 157,000 individuals corresponding to 13,538 haplotypes, at the 1% level.

4. Conclusions

By avoiding the assumptions of theoretical models, the empirical method presented here for the estimation of saturation levels of the number of Y-STR haplotypes is only dependent on the observed variables of sample size and number of haplotypes. Geography was shown to be a minor factor at the European level, supporting the use of large international databases for match searching. Anyway, under our approach, it is predictable that a fivefold increase (up to 60,000 individuals) of the European Y-STR database will reveal about double (11,308) the already sampled haplotypes.

Acknowledgements

This work was partially supported by a research grant to LP (SFRH/BPD/7121/2001) from Fundação para a Ciência e a Tecnologia and IPATIMUP by Programa Operacional Ciência, Tecnologia e Inovação (POCTI), Quadro Comunitário de Apoio III.

References

[1] A. Helgason, S. Sigurðardóttir, J. Gulcher, K. Stefánsson, R.R. Ward, Sampling saturation and the European mtDNA pool: implications for detecting genetic relationships among populations, in: C. Renfrew, K. Boyle (Eds.), Archaeogenetics: DNA and the Population Prehistory of Europe, 2000, pp. 285–294, Cambridge.

[2] H. Pfeiffer, P. Forster, C. Ortmann, B. Brinkmann, The results of an mtDNA study of 1200 inhabitants of a German village in comparison to other Caucasian databases and its relevance for forensic casework, Int. J. Legal Med. 114 (2001) 169–172.