

Luísa Pereira · Carla Cunha · António Amorim

Predicting sampling saturation of mtDNA haplotypes: an application to an enlarged Portuguese database

Received: 27 February 2003 / Accepted: 12 December 2003 / Published online: 11 February 2004
© Springer-Verlag 2004

Abstract An enlarged mtDNA database ($n=549$) for the Portuguese population, comprising HVRI and HVRII regions is reported. This database was used to test the effect of sample size on the estimation of relevant parameters such as haplotype diversity, number of different haplotypes, nucleotide diversity and number of polymorphic positions. Simulations were performed generating sets of random subsamples of variable sizes ($n=50, 100, 200, 300$ and 400). The results show that while haplotype and nucleotide diversities do not vary significantly with sample size, the numbers of haplotypes and polymorphic positions rise continuously inside the tested interval. These trends are interpretable by the evolution of the proportions of sequences that are found once or twice, which drop dramatically as sample size increases, with the corresponding rise in the frequency of those encountered 3 times or more. The generated data were also used to extrapolate saturation curves for the referred parameters. When considering for instance the number of haplotypes, it is shown that a sample size of 1,000 individuals is required for practical saturation (defined as the point where a sample size increase of 100 individuals corresponds to an increment in the diversity measure below 5%). For HVRII the same level is reached at $n=900$ and $n=1,300$ is needed when both regions are analysed simultaneously. Consequently, we can infer that currently used sample sizes are still rather inadequate for both anthropological and forensic purposes.

Electronic Supplementary Material Supplementary material is available in the online version of this article at <http://dx.doi.org/10.1007/s00414-003-0424-1>

Keywords Saturation · mtDNA · Portugal · HVRI · HVRII

Introduction

Statistical problems, namely sampling, are particularly acute when dealing with hypervariable loci such as mtDNA sequences or Y-chromosome STR haplotypes. Among these problems are of notorious practical interest those related to frequency estimations (for instance in evaluating genetic relationships between populations, Helgason et al. 2000) 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, Pfeiffer et al. 1999).

It has been shown by extrapolating from actual European mtDNA HVRI sequence data that current sample sizes are clearly insufficient for adequate population comparisons, requiring a 3 to 4-fold increase of the sampling effort per population (Helgason et al. 2000). In accordance, Pfeiffer et al. (2001) when analysing the same type of data in an unusually large ($n=1,200$) German sample registered that although haplotype diversity values reached a plateau around $n=400$, the number of different haplotypes continued to steadily increase.

In this work we tried to confirm and extend these results, using a different approach and an enlarged mtDNA Portuguese database ($n=549$) including HVRI and HVRII sequences. To avoid the assumptions (in particular the infinite alleles model) underlying the method of Helgason et al. (2000) we performed simulations generating sets of random subsamples of variable size in sufficient numbers for a direct curve-fitting approach allowing the safe extrapolation of saturation levels. Furthermore we have extended the analyses to HVRII data and compared the results from Pfeiffer et al. (2001) on Germans with the Portuguese, already known to display lower haplotype diversity (Pereira et al. 2000).

L. Pereira (✉) · C. Cunha · A. Amorim
IPATIMUP (Instituto de Patologia e Imunologia Molecular da Universidade do Porto), R. Dr. Roberto Frias s/n,
4200-465 Porto, Portugal
Tel.: +351-225570700, Fax: +351-225570799,
e-mail: lpereira@ipatimup.pt

C. Cunha · A. Amorim
Faculdade de Ciências da Universidade do Porto, Portugal

Material and methods

A total of 549 unrelated individuals born in Portugal were analysed for HVRI and HVRII according to conditions described in Pereira et al. (2000). The results are provided as electronic supplementary material. All Portuguese districts were sampled (*VC* Viana do Castelo, *B* Braga, *P* Porto, *VR* Vila Real, *BG* Bragança, *AV* Aveiro, *VI* Viseu, *G* Guarda, *C* Coimbra, *L* Leiria, *CB* Castelo Branco, *S* Santarém, *LX* Lisboa, *SB* Setúbal, *PL* Portalegre, *EV* Évora, *BJ* Beja, *F* Faro; mean of 30.5 individuals/district) and grouped into three large regions: north Portugal (*NP*) with a sample size of 187; central Portugal (*CP*) 239; and south Portugal (*SP*) 123. This division corresponds approximately to the country division by the two main rivers Douro and Tagus. The sequences previously published in Pereira et al. (2000) are marked in the present database with *, and some misassignments to the large Portuguese regions were corrected (and are indicated by **), namely, some samples from the Lisbon district previously classified as south Portuguese are now included in central Portugal. Positions considered for analysis were between 16024–16365 bp, for HVRI and between 073–340 bp for HVRII, according to the reference sequence of Anderson et al. (1981). The hypervariable positions 16182 and 16183 in HVRI and the indels at positions 309 and 315 in HVRII were not considered for analysis (Bendall and Sykes 1995).

To evaluate the sample size effect, sets of 4 subsamples sized 50, 100, 200, 300 and 400 were randomly generated.

Diversity measures were estimated using Arlequin 2.0 software (Schneider et al. 2000).

Curve fittings were obtained through the CurveExpert version 1.37 package (<http://www.ebicom.net/~dhyams/cvxp.htm>), excluding all oscillating functions, even if they were classified as best fitted.

Results and discussion

Sample size effect on standard mtDNA diversity measures using the Portuguese database

The correlations between sample sizes as obtained by the random subsampling method and some standard measures

of diversity for HVRI, HVRII and both regions together, are displayed in Fig. 1.

With respect to the haplotype diversity (Fig. 1A), for HVRI it increases with the sample size, up to a maximum value of 0.964 for a sample of 300 individuals, almost 2% lower than the maximum value observed in Germany, at $n=400$ (Pfeiffer et al. 2001). For HVRII, the maximum (0.907) is attained at $n=400$, while the effect of sample size is indiscernible when both regions are considered.

On the contrary, the number of haplotypes increases monotonously with sample size, for each region as well as for both simultaneously (Fig. 1B).

Nucleotide diversity (Fig. 1C) correlates with sample size in a fashion similar to haplotype diversity, but saturation levels are reached earlier ($n=100$) for HVRI. When HVRI and HVRII are considered simultaneously, an intermediate situation between each of the regions is revealed (showing, in other words, that average HVRI+HVRII nucleotide diversity is systematically lower than the HVRI alone, no matter what the sample size is).

On the contrary, the number of polymorphic positions (Fig. 1D), as observed before for the number of haplotypes, does grow steadily with the increase of sample size, reaching in the total sample 35%, 22% and 29% of the segment length for HVRI, HVRII and both, respectively.

The main reason explaining the observation that incorporating HVRII information does not augment the overall nucleotide diversity (although contributing to increments on number of haplotypes) results from the large proportion of recurrent, highly mutable, HVRII positions (Forster et

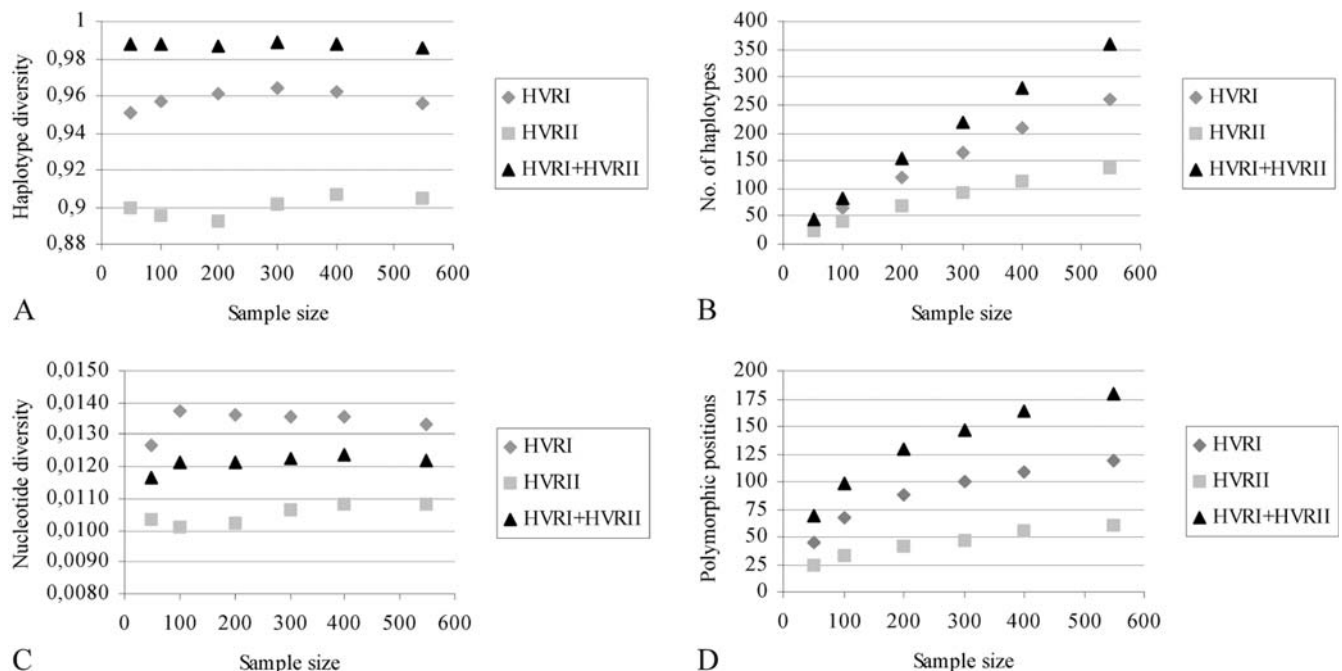


Fig. 1 Correlations between sample size and **A** haplotype diversity, **B** number of haplotypes, **C** nucleotide diversity and **D** number of polymorphic positions, for mtDNA HVRI, HVRII and both regions (HVRI+HVRII)

al. 2002), consequently determining the variation of the two segments to behave as nearly independent. This feature renders HVRII poorly informative for population studies. However, for identification purposes, as in forensic investigations, HVRII can still be very valuable in haplotype discrimination. We must stress that for this analysis we have not considered the highly heteroplasmic nucleotide stretch 303–309, specifically because this property (even samples from the same individual can vary significantly for these positions, and its heteroplasmy can be due to technical artefacts, such as DNA polymerase slippage and not to genetic causes), would significantly increase the diversity for HVRII.

The apparent contradiction between the finding of a saturating value for haplotype diversity and the constantly increasing number of haplotypes is explained by the fact that the probability of selecting at random two individuals with different haplotypes depends on the frequency of haplotypes present more than once in the sample, as already put forward by Helgason et al. (2000).

The proportion of the sample occupied by sequences that are observed once, twice, or more than twice is represented in Fig. 2. For HVRI, the proportion of unique sequences starts with a value around two-thirds ($n=50$), decreasing to about one-third in the total sample ($n=549$). The proportion of sequences represented twice remains almost constant, but the proportion of sequences occurring 3 times or more increases in an inverse way to the unique sequences and becomes predominant (more than one-half for $n=549$). The largest contribution for this last class results from the Cambridge Reference Sequence, CRS (19.9%), reaching nearly double the proportion reported for Germany (10.3%, Pfeiffer et al. 2001). For HVRII, the proportion of unique sequences decreases very fast from an already low value of 39.0% to just 14.9%, so that the proportion of sequences present more than twice represents half of the sample of 50 individuals and three-quarters of the total sample. Considering both regions, the proportion of unique sequences remains above one-half for the total sample.

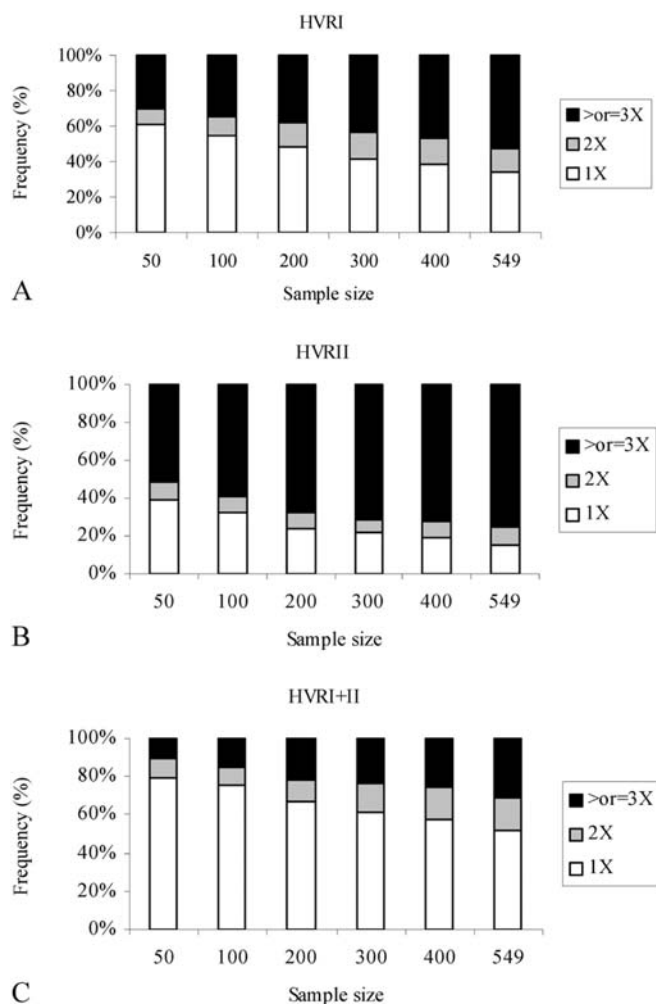


Fig. 2 Variation according to sample sizes of the proportions of sequences found once (1X), twice (2X) or more than twice (>or=3X), for mtDNA regions HVRI **A**, HVRII **B** and both **C**

Theoretical sampling saturation curves

The best fitting function for the correlation between the number of haplotypes and sample size (under the conditions defined in the Material and Methods section and after elimination of those either requiring more than two parameters or failing to converge in one of the HVR regions tested) was obtained with a saturation growth-rate model of the form $y=ax/(b+x)$ as shown in Fig. 3. Defining saturation level as the point for which increments of 100 individuals do not raise the number of haplotypes by more than 5%, the corresponding values are around $n=1,000$ for HVRI (with 370 haplotypes), $n=900$ for HVRII (with 169 haplotypes) and $n=1,300$ for both regions combined (with 638 haplotypes). However, if we establish instead 1% as the satisfactory level, the required sample sizes increase more than 2-fold: 2,800, 2,100 and 3,400, corresponding to numbers of haplotypes of 557, 218 and 987, respectively.

The theoretical curves for the correlation between sample size and the number of polymorphic positions at HVRI, HVRII or both, are shown in Fig. 4 using the same curve fit as above. As expected from the intrinsic limitations of the available sequence length under analysis, they reach saturation (at the 5% level) much earlier: $n=400$ for HVRI (number of polymorphic positions 108), $n=300$ for HVRII (number of polymorphic positions 49) and again 400 for both (number of polymorphic positions 161). At 1%, $n=1,000$ in all cases, with numbers of polymorphic positions being 124, 60 and 184, respectively. This apparently surprising fact, that HVRII does not increase the required sample size, was already explained above: this region shows a limited number of variable, although very polymorphic and recurrently mutating sites, so that database enlargement is not expected to substantially increase the detection of new ones.

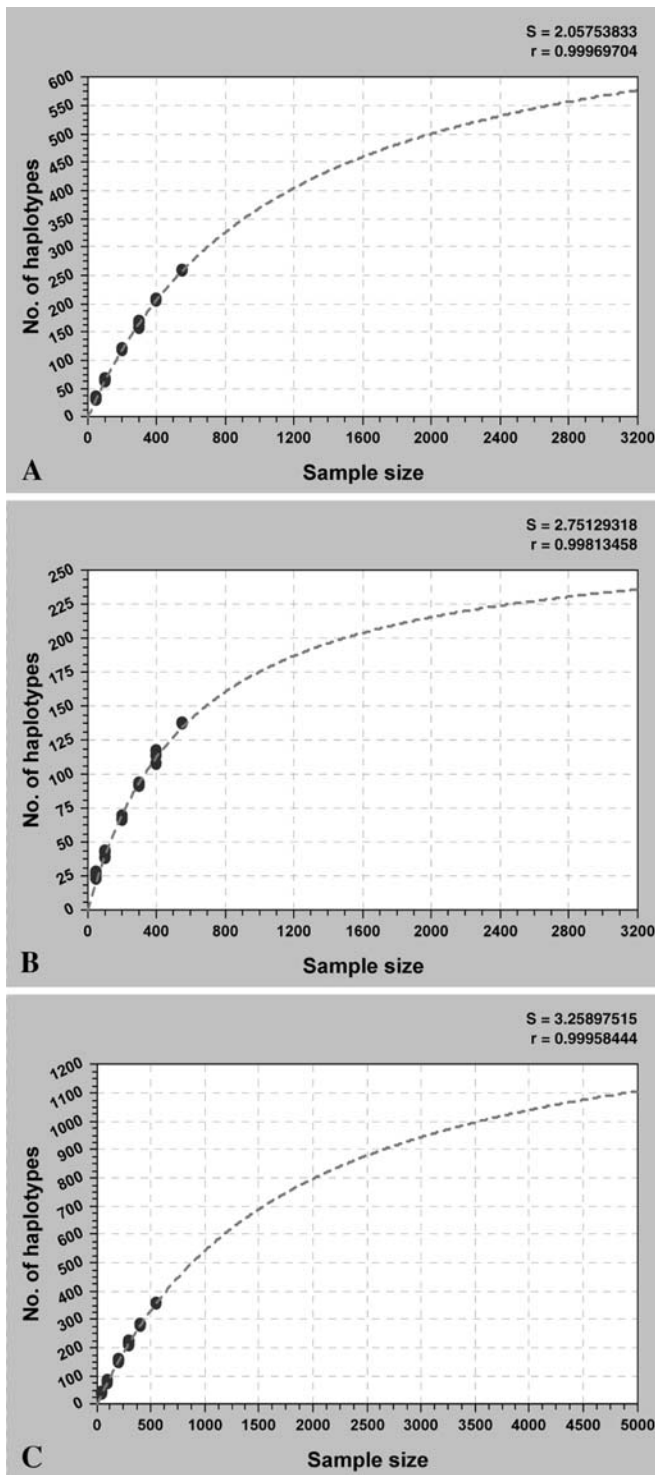


Fig. 3 Sample sizes and number of haplotypes: regression curves for HVRI (A), HVR II (B) and HVRI+HVR II (C). S standard error, r correlation coefficient. All curves are of the form $y=ax/(b+x)$, and coefficients were $a=775.77$ and $b=1098.72$ for HVRI, $a=279.50$ and $b=588.64$ for HVR II, $a=1493.22$ and $b=1742.11$ for HVRI+HVR II

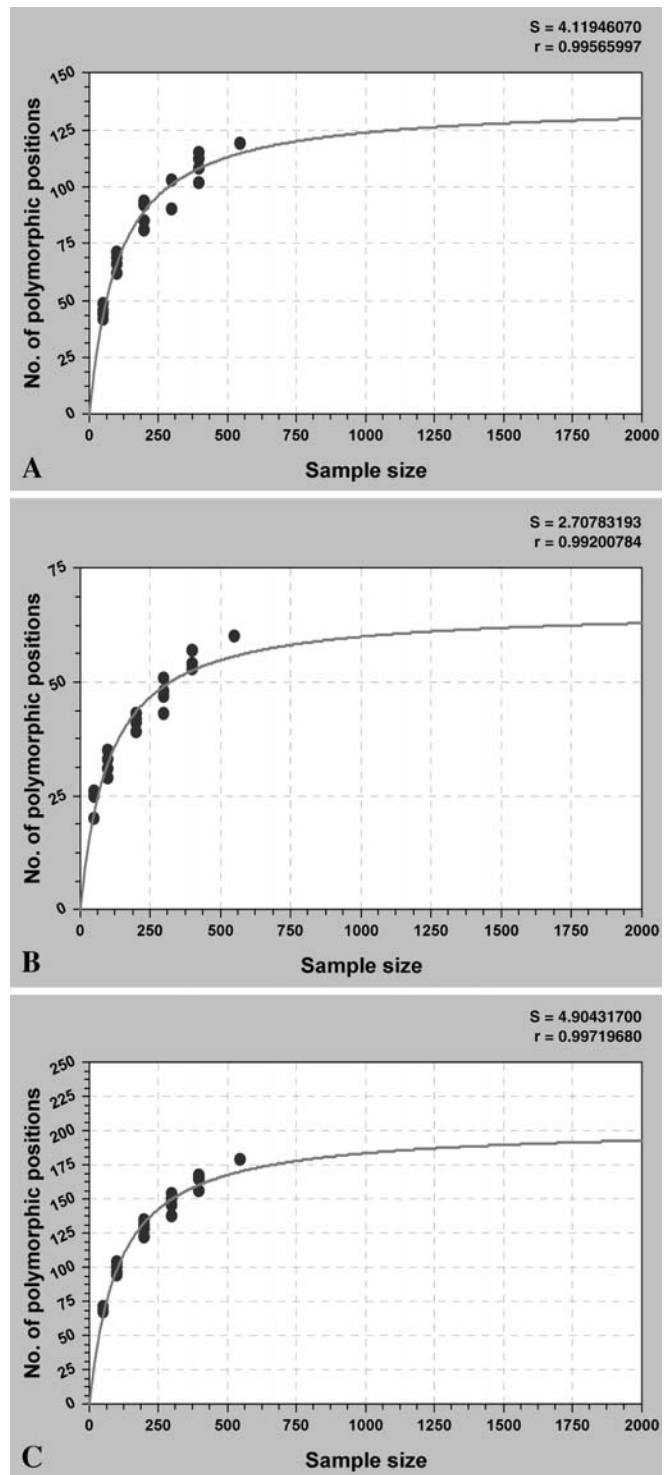


Fig. 4 Sample sizes and number of polymorphic positions for HVRI A, HVR II B and HVRI+HVR II C. S standard error, r =correlation coefficient. All curves are of the form $y=ax/(b+x)$, and coefficients were $a=136.97$ and $b=105.72$ for HVRI, $a=66.22$ and $b=103.63$ for HVR II, $a=203.22$ and $b=105.07$ for HVRI+HVR II

Comparisons with other databases

A previous report (Pereira et al. 2000) has shown that mtDNA diversity of the Portuguese population is low when compared with central European populations. Although containing an unusually high proportion of non-European lineages, those for which the origin is attributable to this continent (the vast majority of the database) are a reduced subset of those found in central Europe. This was confirmed with the enlargement of the database now reported and so it would be expected that sample saturation values for any of the diversity measures in the Portuguese would be lower than the corresponding ones in central Europe. This is indeed the case, if for instance the HVRI data for Germany (Pfeiffer et al. 2001) are used for comparison. For example, the saturation value for haplotype diversity is reached at a smaller sample size ($n=300$) than in Germany ($n=400$).

Focusing on the theoretical saturation curves we obtained for the number of haplotypes, a comparison is possible with the alternative method of Helgason et al. (2000), which assumes the infinite alleles model and was applied to HVRI European data (using however a slightly shorter amplicon and different criteria). While in Portugal a saturation value around $n=1,000$ was found with 430 haplotypes, the corresponding values for Germans would be 1,700 and 440, respectively. These comparisons have however to be taken with caution, since the data cited by Helgason et al. (2000) have been shown to contain several errors (Arnason 2003).

Unfortunately the German data of Pfeiffer et al. (2001) do not allow a reliable curve fitting test, due to a big gap in subsampling intervals (no data for the range $n>350$ to $n=1,200$) and because only the mean values for the subsamplings were reported. However, the curve fitting using the available interval and without duplicates was rather good (standard error=1.3190; correlation coefficient=0.9998), but the prediction of expected number of haplotypes for a sample of 1,200 individuals (370) was not very close to the observed value (460) seemingly (as it can be seen in Fig. 1A of that reference) due to an inflection in the curve for $n>350$.

Conclusions

Our analyses confirm the inferences put forward by Helgason et al. (2000), namely that different populations require varying sampling efforts. The proportion of lineages found once or twice was suggested as a guideline for judging the adequacy of the sample size, and it was advised that sampling should be continued until only no more than 70–80% of the sample belonged to that category. This

reasoning is based on very complex expectations on the genetic structure and demography of the populations under scrutiny. Our results, generated under an assumption-free approach, agree in general, and particularly on the fact that current sampling of large populations demands a substantial increase in numbers.

Indeed, while Helgason et al. (2000) suggested a 3 to 4-fold increase, based on the abovementioned guidelines, our method, which provides a more flexible way of gauging sample efficiency, indicates for the same type of populations at least a 3 to 5-fold increase.

These results seem particularly relevant for population analyses in general (and particularly for the definition of private lineages) and also in forensic cases, where simple matching strategies are currently used, both for individual identification as well as for ethnic or geographical associations.

Acknowledgments We acknowledge Alexandra Lopes for collecting samples from central and south Portugal and providing them for this study. Luísa Pereira has a post-doctoral grant from Fundação para a Ciência e a Tecnologia (SFRH/BPD/7121/2001). IPATIMUP is supported by Programa Operacional Ciência, Tecnologia e Inovação (POCTI), Quadro Comunitário de Apoio III.

References

- Anderson S, Bankier AT, Barrell BG et al. (1981) Sequence and organization of the human mitochondrial genome. *Nature* 290: 457–465
- Arnason E (2003) Genetic heterogeneity of Icelanders. *Ann Hum Genet* 67:5–16
- Bendall KE, Sykes BC (1995) Length heteroplasmy in the first hypervariable segment of the human mtDNA control region. *Am J Hum Genet* 57:248–256
- Forster L, Forster P, Lutz-Bonengel S, Willkomm H, Brinkmann B (2002) Natural radioactivity and human mitochondrial DNA mutations. *Proc Natl Acad Sci U S A* 99:13950–13954
- Helgason A, Sigurðardóttir S, Gulcher J, Stefánsson K, Ward R (2000) Sampling saturation and the European mtDNA pool: implications for detecting genetic relationships among populations. In: Renfrew C, Boyle K (eds) *Archaeogenetics: DNA and the population prehistory of Europe*. McDonald Institute for Archaeological Research, Cambridge, pp 285–294
- Pereira L, Prata MJ, Amorim A (2000) mtDNA diversity in Portugal: not a genetic edge of European variation. *Ann Hum Genet* 64:491–506
- Pfeiffer H, Brinkmann B, Hühne J et al. (1999) Expanding the forensic German mitochondrial DNA control region database: genetic diversity as a function of sample size and microgeography. *Int J Legal Med* 112:291–298
- Pfeiffer H, Forster P, Ortmann C, Brinkmann B (2001) 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* 114:169–172
- Schneider S, Roessli D, Excoffier L (2000) Arlequin ver. 2.000: a software for population genetics data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland