

Table 1**Rank of Icelanders, Danes, and Germans for HVS1 Summary Statistics among 26 European Populations**

POPULATION	RESULTS OF											
	Helgason et al. (2000)						Arnason (2003)					
	<i>N</i>	<i>k</i>	Rank				<i>N</i>	<i>k</i>	Rank			
		Gene Diversity	Pairwise Difference	θ_k	θ_s			Gene Diversity	Pairwise Difference	θ_k	θ_s	
Icelanders	447	125	9	8	14	13	520	128	7–9	9	13	11
Danes	31	25	5	2	13	24	33	19	23	24	25	25
Germans	418	219	6	14	2	1	423	217	10–15	21	2	1

NOTE.—The population with the highest value has a rank of 1, and the population with the lowest value has a rank of 26. *N* is sample size, and *k* is the number of distinct haplotypes. θ_k and θ_s are population-mutation rate parameters that are based on the number of haplotypes and the number of variable sites, respectively.

Address for correspondence and reprints: Dr. Jacek Zaremba, Department of Genetics, Institute of Psychiatry and Neurology, Sobieskiego 9, Warsaw, 02-957, Poland. E-mail: zaremba@ipin.edu.pl © 2003 by The American Society of Human Genetics. All rights reserved.
0002-9297/2003/7304-0024\$15.00

Am. J. Hum. Genet. 73:974–975, 2003

Erroneous Claims about the Impact of Mitochondrial DNA Sequence Database Errors

To the Editor:

In a recent letter to the *Journal*, Herrnstadt et al. (2003, p. 1,585) assert that sequence errors in an mtDNA study of Icelanders contributed to an “erroneous conclusion about the genetic diversity of these people.” Herrnstadt et al. (2003) do not explicitly refer to the study in question, nor do they provide any evidence to support their allegation. Rather, they cite a recent article by Arnason (2003) that states that errors in sequences obtained from a database had a material effect on the results and, hence, on conclusions of an article published in the *Journal* by Helgason et al. (2000). We demonstrate here, contrary to this claim and its reiteration by Forster (2003) and Herrnstadt et al. (2003), that sequence errors have no impact on the conclusions of Helgason et al. (2000) about the genetic diversity of Icelanders.

Helgason et al. (2000) analyzed a total of 4,064 mtDNA control-region sequences: 2,969 from hypervariable segment 1 (HVS1) in 26 populations and 1,095 from hypervariable segment 2 (HVS2) in 10 populations. A subset of these sequences was obtained from HVRBASE (Handt et al. 1998), among which were 140 HVS1 sequences from Denmark and Germany (Richards et al. 1996). Arnason (2003) points out that some of these latter sequences were incorrectly recorded in HVRBASE. After correcting the erroneous sequences, Arnason recalculates genetic diversity statistics presented in table 1

of Helgason et al. (2000) and concludes that “[t]he estimation of statistics and relative rank of countries with respect to diversity and effective population size is materially affected by the errors” (Arnason 2003, p. 9). He moreover concludes that “[c]laims about a special genetic homogeneity of Icelanders relative to European populations would be suspect to the extent that they depended on anomalous data instead of the primary data” (Arnason 2003, p. 14). In an accompanying editorial comment, Forster (2003, p. 2) elaborates: “Arnason demonstrates [that HVRBASE] is riddled with copying errors in the case of the Danes and Germans, resulting in a *qualitatively different* ranking of Icelandic genetic diversity” (emphasis added).

Our comparison of the German and Danish sequences, stored in HVRBASE, with the original sequences submitted to GenBank has revealed that 14 of 33 Danish and 29 of 107 German HVS1 sequences contained database transcription errors. The impact of these errors on the summary statistics (gene diversity and mean pairwise differences, θ_k and θ_s , respectively) calculated by Helgason et al. (2000) for Icelanders, Danes, and Germans can be evaluated through a comparison with Arnason’s (2003) recalculations (see table 1).

Table 1 shows that the rank of the Danish sample changes for all summary statistics, a finding that is not surprising, given that almost half of the Danish sequences contained errors. In the case of the Germans, we observe noticeable changes with respect to gene diversity and mean pairwise differences but no changes in rank for θ_k and θ_s . What is most important in the present context is that table 1 shows that the erroneous Danish and German sequences have almost no effect on the relative position of Icelanders among the 26 European populations for all four summary statistics. This clearly contradicts the aforementioned claims, made by Arnason (2003) and repeated by Forster (2003) and Herrnstadt et al. (2003), that sequence errors in the study by Helgason et al. (2000) have an impact on conclusions about the relative genetic diversity of the Icelanders.

Arnason's (2003) conclusion that Icelanders are among the most genetically heterogeneous populations in Europe was, in fact, entirely based on his choice to interpret gene diversity and mean pairwise differences as the most informative statistics for detecting differences between the mtDNA heterogeneity of Icelanders and other European populations. However, as Helgason et al. (2003) have shown by means of simulation, such pairwise statistics are largely insensitive to the impact of different effective population sizes during periods such as the past 1,100 years (the time since the settlement of Iceland). Population genetic simulations demonstrate that these pairwise statistics show little change on average but considerable increase in variance as effective population size decreases (Helgason et al. 2003). Hence, it is not surprising to observe increased values of gene diversity and mean pairwise differences in populations experiencing high levels of drift due to small effective size. In contrast, Helgason et al. (2003) show that summary statistics that are based on the number of haplotypes or the number of segregating sequence sites (such as θ_k and θ_s) are highly sensitive to differences in effective population size during recent population history. On the basis of simulations and analyses of autosomal, Y-chromosome, and mtDNA data sets, the results of Helgason et al. (2003) strongly indicate that Icelanders are among the most genetically homogeneous European populations and that this homogeneity is the result of a relatively small effective population size.

We agree with Herrnstadt et al. (2003) that more effort should be given to quality control in the sequencing process and database construction. Indeed, as an additional quality-control measure in the Helgason et al. (2000) study, we blindly sequenced and proofread control region sequences from maternal relatives for >25% of the 402 Icelanders included in the study. Consequently, our Icelandic mtDNA sequences are likely to be less affected by errors than most data sets in the literature. However, in today's world of large-scale sequencing projects, er-

rors, whether introduced by the sequencing process itself or by transcription into databases, are likely to be an inescapable fact of life. Bearing this in mind, it is important and productive to carefully assess the impact of sequence errors on scientific results and conclusions. It is less productive to make (Arnason 2003) or uncritically repeat (Forster 2003; Herrnstadt et al. 2003) incorrect claims about the impact of such sequence errors.

AGNAR HELGASON AND KÁRI STEFÁNSSON

deCODE Genetics
Reykjavik

References

- Arnason E (2003) Genetic heterogeneity of Icelanders. *Ann Hum Genet* 67:5–16
- Forster P (2003) To err is human. *Ann Hum Genet* 67:2–4
- Handt O, Meyer S, von Haeseler A (1998) Compilation of human mtDNA control region sequences. *Nucleic Acids Res* 26: 126–129
- Helgason A, Nicholson G, Stefánsson K, Donnelly P (2003) A reassessment of genetic diversity in Icelanders: strong evidence from multiple loci for relative homogeneity caused by genetic drift. *Ann Hum Genet* 67:281–297
- Helgason A, Sigurðardóttir S, Gulcher JR, Ward R, Stefánsson K (2000) mtDNA and the origin of the Icelanders: deciphering signals of recent population history. *Am J Hum Genet* 66: 999–1016
- Herrnstadt C, Preston G, Howell N (2003) Errors, phantom and otherwise, in human mtDNA sequences. *Am J Hum Genet* 72:1585–1586
- Richards M, Côrte-Real H, Forster P, Macaulay V, Wilkinson-Herbots H, Demaine A, Papiha S, Hedges R, Bandelt H-J, Sykes B (1996) Paleolithic and neolithic lineages in the European mitochondrial gene pool. *Am J Hum Genet* 59:185–203 (erratum 59:747)

Address for correspondence and reprints: Dr. Agnar Helgason, deCODE Genetics, Sturlugata 8, 101 Reykjavik, Iceland. E-mail: agnarh@hi.is
© 2003 by The American Society of Human Genetics. All rights reserved.
0002-9297/2003/7304-0025\$15.00