International Congress Series 1288 (2006) 58–60

www.ics-elsevier.com

A compact population analysis test using 32 SNPs with highly diverse allele frequency distributions

C. Phillips [a,*], J. Sanchez [b], M. Fontadevila [a], A. Gómez-Tato [c], J. Alvarez-Dios [c], M. Calaza [c], M. Casares de Cal [c], D. Ballard [d], A. Salas [a], A. Carracedo [a]

The SNPforID Consortium[1]

[a] *CeGen, Molecular Medicine Unit, University Hospital, Santiago de Compostela, Spain*
[b] *Department of Forensic Genetics, University of Copenhagen, Denmark*
[c] *Faculty of Mathematics, University of Santiago de Compostela, Spain*
[d] *Department of Haematology, Queen Mary's School of Medicine, London, UK*

Abstract. A total of 32 population informative SNP markers have been combined into a single multiplex PCR test. Studies of samples from three major population groups indicate strongly contrasting allele frequency distributions in all the loci. A statistical analysis of SNP profiles using measurement of maximum likelihood provides a system for predicting population of origin with a very low error rate. This analysis system will form a core part of an open access web portal. © 2005 Published by Elsevier B.V.

Keywords: SNP; AIM; Population genetics; Forensic analysis

1. Introduction

The ability to predict the population of origin of an individual has been a long-standing aim in forensic genetics that has met with only limited success. Much depends on being able to find the between-population variation that forms only a small proportion (estimated to be 9–13%) of the total variation. Autosomal SNPs offer a good opportunity to detect such variation since they are stable and usually show binary polymorphism: making them more

* Corresponding author.
E-mail address: c.phillips@mac.com (C. Phillips).
[1] www.snpforid.org.

0531-5131/ © 2005 Published by Elsevier B.V.
doi:10.1016/j.ics.2005.10.023

prone than other markers to changes in allele frequency from drift. Furthermore, SNPs creating, or close to, gene variants subject to intense positive selection that postdates population divergence can exhibit sharp changes in allele frequency within a defined geographic region. Well-documented examples include the DARC, LCT and MATP genes. The SNPforID Consortium has developed a single multiplex test that uses SNP markers showing highly contrasting allele frequency distributions in three population groups: African, European and East Asian. Loci were selected from previously collected sets of population-specific SNPs and non-binary SNPs, from published lists of Ancestry Informative Markers (AIMs) and from the genes listed above, known to have been subject to recent, localized selection. Our aim has been to provide a complete package for forensic analysis: optimized primer sets for a single multiplex PCR, a simple detection system based on primer extension, validated allele frequencies and a web-based analysis portal that can suggest a population of origin from a submitted SNP profile using a Bayesian approach to the analysis of detected alleles.

2. Materials and methods

A total of 32 SNPs were combined from much larger sets of collected markers exhibiting very high between-population F_{ST} and minimal within-population F_{ST}. The final selection was made on the basis of a markers ability to differentiate the three population groups analyzed, the multiplex performance and the chromosome location, in that order of preference. The primer design process and conditions for multiplex PCR and single base extension detection were as previously described [1], with 1–10 ng of target DNA routinely used for amplification.

The statistical analysis to allow a suggested population of origin calculated the maximum likelihood of the profile belonging to a particular population from the allele combinations detected, given by the probability:

$$P((x_1,\ldots,x_m)|\mathrm{Ind}\in\mathrm{pop}_i) = P(x_i|\mathrm{Ind}\in\mathrm{pop}_i)\ldots P(x_m|\mathrm{Ind}\in\mathrm{pop}_i)$$

where $x_1,\ldots,x_m$ denotes allele frequencies, Ind the profile and pop_i is each population.

This assumes random variable independence, i.e. the individual allele frequency estimates for each population can be combined into a joint distribution, with no association between markers in the group. This underlines the need for the SNP set to be freely assorting. The allele frequency distributions were estimated using a training set of 90 samples from each population group (Mozambican, Galician and Taiwanese). Zero observations were substituted by the frequency: $1/(2n+1)$ where n = samples. The theoretical error rate of the system was estimated using re-classification of the training set profiles, cross-validation and bootstrapping (1000 replications). In addition, 60 new profiles from different populations in each group (Somali, Danish and Chinese, respectively) were tested using the original training set. The analysis system was designed from the start to be able to assess partial profiles.

3. Results

The allele frequency estimate distributions of the 32 SNPs from each training set population are summarized in Fig. 1. The error rate for classifications using the three

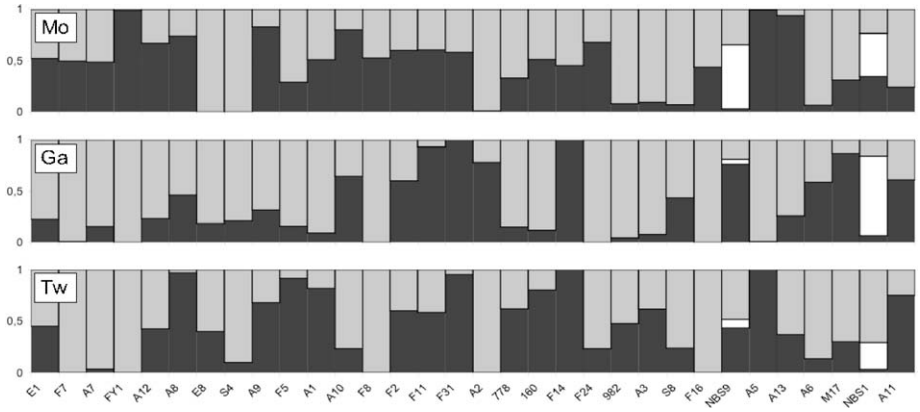

Fig. 1. Allele frequency distributions observed for 32 SNP markers in the three training set population samples. Mo: Mozambicans, Ga. Galicians, Tw: Taiwanese. Light grey bars represent the HGMP reference sequence bases. White bars represent the third allele in the two non-binary SNPs: NBS1 and NBS9.

methods of assessment was 0% in all cases except for Europeans and East Asians that gave a 0.01% error rate from bootstrapping for both comparisons. An error rate of 0.16% was observed for the classification of novel profiles (1 of 603 samples assessed to date).

4. Discussion

By selecting the best possible SNP markers in terms of between-population F_{ST}, we have been able to develop a single multiplex test that generates a profile highly indicative of ancestry from one of three population groups. The Bayesian approach developed for the statistical assessment of profiles that can provide a suggested population of origin, exhibits a very low theoretical error rate. Amongst factors that could lead to a higher error rate in practice are the possibility of population admixture and encountering individuals with immediate co-ancestry (i.e. parental or grandparental). We are currently refining the analysis system to detect admixture at both the population and the individual level in order to prevent the erroneous classification of such profiles.

Acknowledgement

The authors would like to acknowledge the facilities for high-throughput SNP genotyping provided by The Spanish National Genotyping Centre, CeGen.

Reference

[1] J.J. Sanchez, et al., Development of a multiplex PCR assay with 52 autosomal SNPs, Int. Congr. Ser. 1288 (2006) 67–69 (this volume).