FREQUENCY MULTIVARIATE ANALYSIS OF LDLR, GYPA, HBGG, D7S8 AND GC IN 12 DIFFERENT POPULATIONS.

E. Arroyo, C. Asperilla, L. Prieto, M. Herrera, J. M. Ruiz de la Cuesta.

Departamento de Toxicología y Legislación Sanitaria. Escuela de Medicina Legal. Facultad de Medicina. Universidad Complutense de Madrid. 28040-Madrid. Spain.

INTRODUCTION

The object of this work is to establish comparisons among those populations studied for LDLR, GYPA, HBGG, D7S8 and GC loci, all of them contained in the "Amplitype PCR PM Amplification and Typing Kit". As far as we know, just twelve populations have been studied for the whole set of markers. Other authors have carried out comparisons among different populations tested for the same set of loci, but none have established a multivariate relation among the series. In this paper, a twofold multivariate analysis was tried: first, a Dendrogram Analysis (DA) and then a Principal Component Analysis (PCA).

MATERIAL AND METHODS

First of all, a matrix was constructed with the gene frequencies cited in the literature plus a sample studied in the Forensic Biology Laboratory of the Legal Medicine School of the Complutense University (207 apparently health non related individuals living in Madrid , Spain).

The set of samples obtained can not be considered a worldwide distribution of frequencies, since it is a mainly caucasian set of samples. However, other racial groups could also be included. The gene frequency matrix was analyzed with the SPSS/PC+ package for microcomputers. A genetic distance matrix was obtained from the gene frequencies matrix and then, a dendrogram was constructed with the former matrix. To calculate the genetic distances between populations a cosine transformation, similar to that proposed by Cavalli-Sforza and Edwards (1967), was used. The UPGMA algorithm was used to construct the dendrogram. Then, to extract the Principal Components (PC), Kaiser-Meyer-Olkin measure of sampling adequacy to the PC model was used and PC's were extracted and plotted. Both multivariate analyses allow to evaluate graphically the genetic proximity of the populations considered.

RESULTS AND DISCUSSION

In the figure 1 a UPGMA dendrogram for the eleven series can be seen. Caucasian samples appear together in the same cluster and at a very short distance: Swiss and Bavarian populations and two series of U.S. caucasians are clustered together with a sample of Hispanics from the south-east of the U.S.A. Spanish series (Galicia and Madrid), plus a sample of American caucasians, are also grouped together and very close to the former caucasian cluster. The other Hispanic samples remain

```
                       Rescaled Distance Cluster Combine

  C A S E   List of 0         5        10        15        20        25
Label         Ref.  +---------+---------+---------+---------+---------+

Hisp-SE-USA     2    -+
Swiss           3    -¦
Cauc-USA(2)     2    -+-+
Bavarians       4    -+ +-----+
Cauc-USA(1)     5    -+ ¦     ¦
Madrid          6    -+-+     +---------------+
Galicia         7    -+       ¦               ¦
Hispanics-USA   5    ---------+               +-----------------------+
Hisp-SW-USA     2    ---+                     ¦                       ¦
Chinese         8    -------------------------+                       ¦
Blacks-USA(1)   5    -------------------------------------------------+
Blacks-USA(2)   2    -+
```

Figure 1. Dendrogram using Average Linkage (Between Groups). UPGMA.
(2) Budowle et al., 1992. (3) Hochmeister MN et al., 1994. (4) Hausmann R et al., 1995. (5) Perkin Elmer. Manufacturer's protocol, 1994. (6) Herrera M., 1995, (7) Pestoni C. , 1995. (8) Huang and Budowle, 1995.

together and, finally, mongoloid and negroid populations - Chinese plus the two U.S. black samples - cluster separately of the caucasian and Hispanic groups. Black samples are at a very short distance between each other and far away from the rest of populations. In conclusion, the DA reflects very well a three major racial groups model.

```
       PLOT OF FACT2 WITH FACT1 BY POPULATION
      +--------------------------------------+
R     ¦                                      ¦
E 1.25¦* Blacks-USA2     Galicia *           +
G     ¦                                      ¦
R     ¦             Caucasian-USA1 *         ¦
      ¦ * Blacks-USA1     Madrid *           ¦
F     ¦                     Swiss *          ¦
A    0¦    Hisp-Southeast-USA *              +
C     ¦               Bavarians *    *       ¦
T     ¦                       Caucasian-USA  ¦
O     ¦                           *          ¦
R     ¦                      Hisp.-USA       ¦
 -1.25¦                                    * +
S     ¦                  Hisp.-Southwest-USA ¦
C     ¦                                      ¦
O     ¦                                      ¦
R     ¦ * Chinese        <- FACTOR 1 ->      ¦
      +--------------------------------------+
         -1.2        -.4        .4       1.2
        -1.6       -.8         0        .8
```

Figure 2. Plot of the two first Principal Components (82.8% of the total variance).

In the case of PC analysis, the gene frequency matrix was evaluated for Kaiser-Meyer-Olkin statistic (0.47105). The result showed a not very good sampling adequacy. However, the initial set of variables could be reduced to three PC's, the first of them explaining 47.4% of the total variance of the data, the second 35.3% and the third 13.5%. The three PC's explain the 96.3% of the total variance of the data.

In the figure 2, a plot of the two first PC's (82.8% of the total variance) can be seen. The results meets quite well the conclusions of the DA: a group of caucasoid populations are very close to each other. Two Hispanic samples plot at the same distance of the caucasian group, and U.S. black samples appear separate of the other populations, the Chinese sample being the most separate of the set. Some subjetive interpretations of the PC plot can be avoided with the help of the DA which shows clearly the grouping of the series based on the genetic distance matrix. Finally, we can list the three following conclusions:

- A twofold multivariate approach can resolve some subjetivity, mainly in the case of PC plotting.
- The combination of DA and PCA seems to support a three major racial groups model for the populations considered in the study.
- The "Amplitype PCR PM Amplification and Typing Kit" is a usefull tool for genetic comparison of populations.

However, it would be better to carry out more studies in order to check the adequacy of the markers contained in the kit for population genetics studies.

REFERENCES

Budowle B, Lindsey JA, DeCou JA, Koos BW, Giusti AM, Comey CT (1995) Validation and population studies of thhe loci LDLR, GYPA, HBGG, D7S8 and GC (PM loci) and HLA-DQ alpha using a multiplex amplification and typing procedure. J. Forensic Sci. 40(1): 45.

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetics analysis: models and estimation procedures. Am. J. Hum. Genet. 19: 233-257.

Hausmann R, HAntschel M, Loetterle J (1995) Frequencies of the 5 PCR-based genetic markers LDLR, GYPA, HBGG, D7S8 and GC in a North Bavarian population. Int. J. Legal Med. 107(4): 227.

Herrera M (1995) Estudio del "Amplitype Polymarker PCR Amplification and Typing Kit". Doctoral Thesis. Universidad Complutense de Madrid.

Hochmeister MN, Budowle B, Borer UV, Dirnhofer R (1994) Swiss population data on the loci HLA-DQA1, LDLR, GYPA, HBGG, D7S8, GC and D1S80. Forensic Sci. Int. 67: 175-184.

Huang NE, Budowle B (1995) Chinese population data on the PCR-based loci HLA-DQ Alpha, Low-Density-Lipoprotein Receptor, Glycophorin A, Hemoglobin γ^G, D7S8, and Group-Specific Component. Hum. Hered. 45: 34-40.

Pestoni C, Muñoz JL, Rodríguez-Calvo MS, Bellas A, García Rivero A, Lareu MV, Barros F, Carracedo A (1995) Actas VII Jornadas Soc. Esp. Med. Legal y Forense. Lérida 4-6 Mayo. p. 473-481.

Perkin Elmer Corp., USA (1994). Amplitype PCR PM Amplifying and Typing Kit Manual. Part. No. N808-0057. p. 3.