Updated version (June 2006, 20th EFI meeting in Oslo) of the teaching session “Biostatistics” presented at the 16th European Histocompatibility Conference, Strasbourg, March 19-22, 2002

HLA data analysis in anthropology: basic theory and practice

Alicia Sanchez-Mazas and José Manuel Nunes

Laboratory of Anthropology, Genetics and Peopling history (AGP), Department of Anthropology and Ecology, University of Geneva (http://agp.unige.ch)

1. Aims

The aim of this communication is to provide immunogeneticists with a basic comprehension of some statistical methods used to analyse HLA data in human populations, and to explain how using some computer programs developed in this scope.

2. Basic theory

2.1. Data

Our interest is here to investigate HLA genetic diversity within and between populations. From an anthropological point of view, the term “population” is most often based on geographical and cultural criteria. However, for medical purposes, for example, one can also consider a “population of patients” by contrast to a “population of healthy people” within a same general population. Some statistical methods used to compare populations from different geographic locations or cultural backgrounds are then also applicable to disease-association studies.

It often occurs that the terms “population” and “sample” are used indifferently. These two terms yet describe completely different entities, and it is crucial to make a distinction between them. By “sample”, we mean a group of individuals chosen from a defined population. Such a sample will be used to *estimate the unknown parameters of the population*, like gene frequencies, as the latter *will never be known*. As an example, we may miss an allele actually present in the population just because the sample-size is limited. The sampling procedure is then decisive to get representative samples and then accurate estimates of the population parameters: who will be sampled, and how many individuals? The answer to the first question will depend on the scope of the study, but some specific information, like the presence of close relatives (completely “unrelated” individuals often do not exist in small populations!) or the fact that the sample consists of blood donors is, in any case, highly instructive. We give in Table 1 the most important information one should gather for a population and the sampled individuals in an anthropological perspective.

Table 1

Population information
- Origin of samples (field study, blood bank, etc)
- Name(s) of the population
- Precise geographic location
- Language, linguistic family
- Social structure and religion
- Demographic information
- Prevalence of specific diseases
- Historical information
- Relationships with other populations
Individual information
- Sampling number and date
- Place of residence
- Individual, maternal and paternal languages
- Declared ethnic and/or linguistic origin
- Sex
- Individual birthdate and birthplace
- Parents and grandparents birth date and place
- Pedigree relationships
- Known diseases

The question of *sample-size* is also of highest importance, and all the more because the number of possible alleles has considerably increased with high-resolution typing techniques. At HLA loci, most populations exhibit a high number of low-frequency alleles, and thus the whole frequency distribution may be inaccurately estimated due to small sample-size. In Table 2 are given the “chances” to miss an allele according both to its frequency in the population (F) and the size of the sample used (N). Obviously, in a sample of 50 individuals, only the most frequent alleles are properly represented.

Table 2

	N = 50	**N = 75**	**N = 100**	**N = 150**	**N = 200**
F = 0.010	0.3660	0.2214	0.1340	0.0490	0.0179
F = 0.020	0.1326	0.0483	0.0176	0.0023	0.0003
F = 0.030	0.0475	0.0104	0.0023	0.0001	0.00001
F = 0.050	0.0059	0.0005	0.00004	0.0000	0.0000

Moreover, a very important sample-size effect is reflected in the results of the statistical tests applied to the data. Depending on the test used, *we may get either false significant or false non-significant results* (these are called type I and type II errors, respectively). For example, we may often wrongly conclude that a given allele is more frequent in patients of a given disease than in controls, only because of a low sample-size. A frequently asked question is then: what is the minimal sample-size for appropriate statistical analyses of HLA data? There is no simple answer. If addressing a precise question, then there are techniques to determine a minimal sample size. For broad purposes, based on our experience and on Table 2, we think that *100 individuals could and should be taken as a very minimal threshold*. However, even under this condition, we ought to be care at any rate of the conclusions we draw from our results.

2.2. Estimating allele frequencies

Even with high-resolution DNA typing techniques, HLA genotypes cannot be defined with certainty at a given locus, due to the fact that hidden alleles (usually called "blank") may still be present: when one unique allele is observed in an individual, the latter may be either homozygous for this allele, or heterozygous for this allele and an unknown one. For that reason, we should always talk about HLA *phenotypes* and not *genotypes*. The putative presence of a blank prevents us from estimating allele frequencies by direct counting of genes (counting once an allele in a heterozygote, and twice in a homozygote): doing that would lead to overestimate the frequencies of the most frequent alleles, and to ignore the fact that undetermined alleles may still be present in the population. In fact, each HLA locus behaves like a system where all alleles are codominant but one (the blank, which can be considered as a recessive allele). A comparable although simpler situation is found for the classical ABO blood group polymorphism. To overcome this difficulty, Bernstein derived in 1924 an expression (known as "Bernstein formula") to estimate allele frequencies from "phenotypic" frequencies in the presence of a recessive. But, while this formula is easily applied with a hand calculator, it only provides a crude idea on gene frequencies. A much more efficient methodology has been developed since then, namely maximum-likelihood and EM (expectation-maximization, also called gene counting) methods, allowing to approach highly probable frequency distributions by an iterative procedure. The principle is that the iterative process stops when the allele frequency or the likelihood differences, between two consecutive iterations, are less than a given threshold. One important thing (unfortunately too often neglected) is that *the basic assumption of either Bernstein formula or maximum-likelihood and EM methods is Hardy-Weinberg equilibrium of the population under study*. This equilibrium has thus to be checked in parallel to any gene frequency estimation, the estimated frequencies being not meaningful otherwise.

2.3. Testing Hardy-Weinberg equilibrium

Under this assumption, allele and genotype frequencies are linked by the simple equations $P_{A_iA_i} = p_i^2$, for a (true) homozygote, and $P_{A_iA_j} = 2p_ip_j$ (i≠j), for a heterozygote. For a system with no recessives, the best way to test this hypothesis is by an exact test (like Fisher's). Unfortunately, *an exact test cannot be applied to current HLA data, due to the putative presence of a blank allele*. Only classical goodness-of-fit methods (like chi-square or G-tests) are allowed. On the other hand, classical goodness-of-fit methods are limited by the fact that they cannot be applied with confidence when the phenotypic distributions include classes with low-frequency expected values (i.e. Np_i^2, for homozygotes, and $2Np_ip_j$, for heterozygotes). In general, one takes a minimal threshold of 5 individuals for such classes. A traditional way to overcome this difficulty is to group low-frequency classes (corrections like

Yate's do not apply in this case). However, the problem with HLA data is that the number of such classes is so high that no criteria for collapsing in one way or another is satisfactory. Actually, there is no ideal solution. However, a reasonable conclusion to draw from classical goodness-of-fit tests is to accept the null hypothesis of no disequilibrium if the final P-value is higher than a given significance level (usually 1% or 5%), and to reject it if the P-value is lower than the significance level *when classes with expected numbers lower than 5 are ignored*. If a P-value is significant only because low-frequency classes raise the total $\chi 2$ or G statistics to a significant value, we can only state that *the conclusion cannot be determined*. If the final conclusion is a significant deviation from Hardy-Weinberg equilibrium, one should not use the estimated allele frequencies without understanding the causes of such a deviation (e.g. heterogeneous sample, mixed, subdivided, or highly endogamous population, strong selective effects on the locus under study, and so on).

Tip: frequently asked statistical questions

What is a P-value?
« it is a measure of the consistency of the observed data with a null hypothesis Ho »

What is a null hypothesis Ho?
« it is a hypothesis stating the equality between observed and expected values, or between two observations »

What is a significant result?
« it is a result suggesting the inconsistency of the observed data with a null hypothesis »

How to evaluate this significance?
« by comparing the P-value to a chosen significance level α »

What is a significance level α?
« it is a level of error accepted when drawing a statistical conclusion »

2.4. Estimating haplotype frequencies

Both the putative presence of a blank allele at each locus and the fact that the gametic phase is rarely known in the sampled individuals render the estimation of haplotype frequencies very tricky. Maximum likelihood methods generally provide accurate frequencies, given that the sample-size is not too small. With HLA data, the number of observed haplotypes is usually so high that the maximum frequency seldom reaches 5%. Here, an important fact to recall is that *the estimation procedure never gives the true haplotype list present in the population, but only a probable one*. Low-frequency haplotypes, although listed, may in fact not be present! A wise assessment, before claiming the occurrence of a given haplotype in a population, is then to verify that the 95% confidence interval of its frequency (approximately the frequency ± two standard deviations) does not contain zero. Otherwise one cannot state that the frequency of this haplotype is different from zero. Even a single copy of this haplotype may not be represented in the sample.

2.5. Testing linkage disequilibrium

Linkage disequilibrium is most often estimated at the individual haplotype level as the non-random association of alleles taken at two different loci. However, the usual linkage disequilibrium coefficient, $D = p_{ij} - p_i . p_j$, is strongly dependent upon allele frequencies and does not take 0 and 1 as minimal and maximal values, respectively. *D coefficients are thus neither comparable among different haplotypes, nor among identical haplotypes in different populations.* For such purposes, the standardised coefficient D' should be used instead. The usual D coefficient can yet be tested for statistical significance (null hypothesis: D=0). This is of the main importance, because even a very frequent haplotype would have no meaning in terms of a particular association among the corresponding alleles if it were found in non-significant linkage disequilibrium (this often occurs, for example, between DPB1 and DRB1 or DQB1 loci). The usual chi-square test can be applied to assess D significance, and here Yate's correction may be useful. However, one should keep a good sense regarding the conclusions, as low-frequency haplotypes can hardly be considered in significant linkage disequilibrium (think, for example, that a 1% frequency in a sample of 100 individuals means that the haplotype is only observed twice).

An alternative way of assessing non-random associations between the alleles of two loci is by using Fisher’s exact test on 2x2 contingency tables. This test is very powerful and applies to low-frequency classes. On the other hand, this approach does not test the same thing as before. It examines *whether the co-occurrence of two given allelic products (or antigens), in the sampled individuals, is higher than by chance,* but a significant result does not mean automatically that the corresponding alleles are on the same chromosome: trans associations are also taken into account. This is like testing allelic association at the phenotypic level, and may be useful for finding donor-recipient compatibilities in organ transplantation.

Finally, besides investigating linkage disequilibrium at each individual two-locus haplotype, one may wonder whether two given loci globally exhibit a statistically significant association, or whether they can be considered as independent. The tests may be carried out as described above, but for the total set of expected haplotypes. In such a case where multiple tests are done, but a global question is asked, up to 5% of significant values may be considered as being due by chance alone if we choose a significance level of 5%. To be sure that the global test is significant, the number of significant values must then be counted after Bonferroni’s correction, which consists, approximately, in dividing the significance level by the number of tests done. The observation of at least one significant haplotype in linkage disequilibrium would then mean a significant global linkage disequilibrium between the two loci. This kind of approach may be applied both to classical goodness-of-fit (chi-square or G-test) and to exact tests assessing allelic association at the haplotypic and genotypic level, respectively.

Tip: Bonferroni’s correction must be applied when many independent tests are performed simultaneously.

The new level of significance (α’ instead of α) is given by

$$\alpha' = 1-(1-\alpha)^{\frac{1}{k}}$$

which is approximately similar to

$$\alpha' = \frac{\alpha}{k}$$

for a total of k tests.

2.6. Testing disease associations

Although it is not the central topic of this communication, the association of a given HLA allele with a particular disease may be assessed by similar approaches as those used for allelic associations. In that case, the best solution is to apply an exact test to a 2x2 contingency table including the number of individuals carrying the allele against those not carrying it among both patient and control population samples. The problem of the blank allele is here not found. However, *one should first define the precise question* (or, in other terms, *the null hypothesis*). One may want to know whether a given allele is associated to the disease. Then an individual 2x2 table is constructed and tested (by Fisher or chi-square with or without Yate’s correction), and the observed P-value is compared to a given significance level (1% or 5%). Alternatively, one may want to verify a global association between a locus and a disease. Many alleles are then tested simultaneously, and Bonferroni’s correction for multiple tests is required.

2.7. Comparing frequencies among populations

A frequent task is to check whether two (or more) particular populations differ genetically from each other. As mentioned in the introduction, these populations may be of different origin or may represent patients and controls for a given disease. While being so common, this task is yet not so easily carried out with available HLA data. Indeed, *one should compare phenotypic or genotypic distributions between populations. Comparing populations using allelic frequencies can only be done when Hardy-Weinberg equilibrium holds.* In both cases, goodness-of-fit tests can be applied by taking into account the sizes of the samples to work on absolute frequencies. The usual problem related to low expected counts still persists.

Another way to test for population differentiation is by computing the *Fst fixation index*, which is proportional to the gene frequency variance among two or more populations. In the field of population genetics, this measure is commonly used as a genetic distance. A powerful procedure to test the significance of this index is provided by a non-parametric permutation approach. After a high number of rounds where individual genotypes are randomly permuted among populations, a null distribution is obtained for Fst, to which the observed Fst is finally compared at a given significance level. This type of procedure has the advantage of being free of many basic assumptions like normality or equality of variance among populations.

3. Basic practice and new developments

Detailed applications of these analyses on concrete examples involving HLA data are given in several documents available at http://geneva.unige.ch/doc/. Most important is the development of procedures for the estimation of allele and haplotype frequencies allowing to include HLA ambiguous data. Useful programs are available under the name Gene[rate] at http://geneva.unige.ch/generate/.

Acknowledgments

We thank the Swiss National Science Foundation (SNF grant #3100A0-112651 to A.S.-M).

Useful references

Nunes, JM (in press) Tools for efficient HLA data handlings. In: Hansen J, Dupont B (eds). HLA 2004. Immunobiology of the Human MHC. Proceedings of the 13th International Histocompatibility Workshop and Conference. Seattle, WA: IHWG Press.

Schneider, S.,Roessli, D., Excoffier, L. (2000) Arlequin: A software for population genetics data analysis. Ver 2.000. Dept. of Anthropology and Ecology, University of Geneva.

Sokal R.Rr, Rohlf F.J. (1995) Biometry: the principle and practice of statistics in biological research. WH Freeman, New York.

Warrens, A.N. (2000) Appendix: statistical considerations in analysing HLA and disease associations. In: Lechler R., Warrens, A. (eds) HLA in health and disease. Academic Press, London.

Weir B.S. (1996) Genetic data analysis II. Sinauer Associates, Sunderland, Massachussets